Chapter 1

Asset Returns

The primary goal of investing in a financial market is to make profits without taking excessive risks. Most common investments involve purchasing financial assets such as stocks, bonds or bank deposits, and holding them for certain periods. Positive revenue is generated if the price of a holding asset at the end of holding period is higher than that at the time of purchase (for the time being we ignore transaction charges). Obviously the size of the revenue depends on three factors: (i) the initial capital (i.e. the number of assets purchased), (ii) the length of holding period, and (iii) the changes of the asset price over the holding period. A successful investment pursues the maximum revenue with a given initial capital, which may be measured explicitly in terms of the so-called *return*. A return is a percentage defined as the change of price expressed as a fraction of the initial price. It turns out that asset returns exhibit more attractive statistical properties than asset prices themselves. Therefore it also makes more statistical sense to analyze return data rather than price series.

1.1 Returns

Let P_t denote the price of an asset at time t. First we introduce various definitions for the returns for the asset.

1.1.1 One-period simple returns and gross returns

Holding an asset from time $t-1$ to t, the value of the asset changes from P_{t-1} to P_t. Assuming that no dividends paid are over the period. Then the *one-period simple return* is defined as

$$R_t = (P_t - P_{t-1})/P_{t-1}. \tag{1.1}$$

It is the profit rate of holding the asset from time $t-1$ to t. Often we write $R_t = 100R_t\%$, as $100R_t$ is the percentage of the gain with respect to the initial capital P_{t-1}. This is particularly useful when the time unit is small (such as a day or an hour); in such cases R_t typically takes very small values. The returns for less risky assets such as bonds can be even smaller in a short period and are often quoted in *basis points*, which is $10,000R_t$.

The *one period gross return* is defined as $P_t/P_{t-1} = R_t + 1$. It is the ratio of the new market value at the end of the holding period over the initial market value.

1.1.2 Multiperiod returns

The holding period for an investment may be more than one time unit. For any integer $k \geqslant 1$, the returns for over k periods may be defined in a similar manner. For example, the *k-period simple return* from time $t-k$ to t is

$$R_t(k) = (P_t - P_{t-k})/P_{t-k},$$

and the *k-period gross return* is $P_t/P_{t-k} = R_t(k) + 1$. It is easy to see that the multiperiod returns may be expressed in terms of one-period returns as follows:

$$\frac{P_t}{P_{t-k}} = \frac{P_t}{P_{t-1}} \frac{P_{t-1}}{P_{t-2}} \cdots \frac{P_{t-k+1}}{P_{t-k}}, \tag{1.2}$$

$$R_t(k) = \frac{P_t}{P_{t-k}} - 1 = (R_t + 1)(R_{t-1} + 1) \cdots (R_{t-k+1} + 1) - 1. \tag{1.3}$$

If all one-period returns $R_t, \cdots, R_{t-k+1}$ are small, (1.3) implies an approximation

$$R_t(k) \approx R_t + R_{t-1} + \cdots + R_{t-k+1}. \tag{1.4}$$

This is a useful approximation when the time unit is small (such as a day, an hour or a minute).

1.1.3 Log returns and continuously compounding

In addition to the simple return R_t, the commonly used *one period log return* is defined as

$$r_t = \log P_t - \log P_{t-1} = \log(P_t/P_{t-1}) = \log(1 + R_t). \tag{1.5}$$

Note that a log return is the logarithm (with the natural base) of a gross return and $\log P_t$ is called the log price. One immediate convenience in using log returns is that the additivity in multiperiod log returns, i.e. the *k period log return* $r_t(k) \equiv \log(P_t/P_{t-k})$ is the sum of the k one-period log returns:

$$r_t(k) = r_t + r_{t-1} + \cdots + r_{t-k+1}. \tag{1.6}$$

An investment at time $t-k$ with initial capital A yields at time t the capital

$$A \exp\{r_t(k)\} = A \exp(r_t + r_{t-1} + \cdots + r_{t-k+1}) = A e^{k\bar{r}},$$

where $\bar{r} = (r_t + r_{t-1} + \cdots + r_{t-k+1})/k$ is the average one-period log returns. In this book *returns refer to log returns* unless specified otherwise.

Note that the identity (1.6) is in contrast with the approximation (1.4) which is only valid when the time unit is small. Indeed when the values are small, the two returns are approximately the same:

$$r_t = \log(1 + R_t) \approx R_t.$$

However, $r_t < R_t$. Figure 1.1 plots the log returns against the simple returns for the Apple Inc share prices in the period of January 1985 – February 2011. The returns are calculated based on the daily close prices for the three holding periods: a day, a week and a month. The figure shows that the two definitions result almost the same daily returns, especially for those with the values between -0.2 and 0.2. However when the holding period increases to a week or a month, the discrepancy between the two definitions is more apparent with a simple return always greater than the corresponding log return.

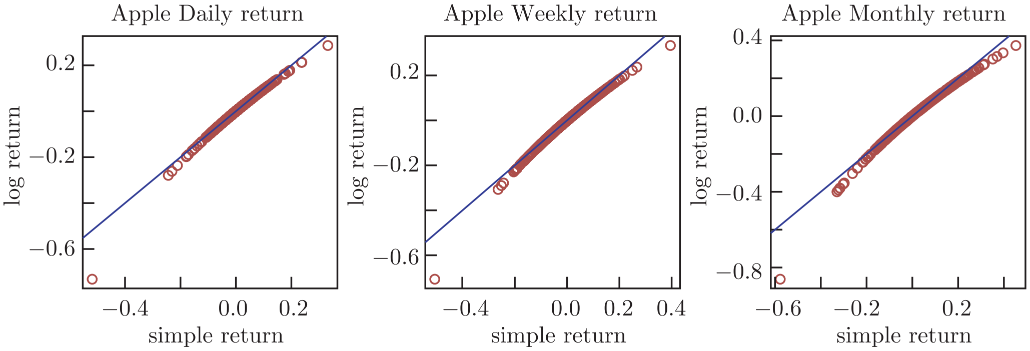


Figure 1.1 Plots of log returns against simple returns of the Apple Inc share prices in January 1985 – February 2011. The blue straight lines mark the positions where the two returns are identical.

The log return r_t is also called *continuously compounded return* due to its close link with the concept of compound rates or interest rates . For a bank deposit account, the quoted interest rate often refers to as ‘simple interest’. For example, an interest rate of 5% payable every six months will be quoted as a simple interest of 10% per annum in the market. However if an account with the initial capital $1 is held for 12 months and interest rate remains unchanged, it follows from (1.2) that the gross return for the two periods is

$$1 \times (1 + 0.05)^2 = 1.1025,$$

i.e. the annual simple return is $1.1025 - 1 = 10.25\%$, which is called the *compound* *return* and is greater than the quoted annual rate of 10%. This is due to the earning from 'interest-on-interest' in the second six-month period.

Now suppose that the quoted simple interest rate per annum is r and is unchanged, and the earnings are paid more frequently, say, m times per annum (at the rate r/m each time of course). For example, the account holder is paid every quarter when $m = 4$, every month when $m = 12$, and every day when $m = 365$. Suppose m continues to increase, and the earnings are paid continuously eventually. Then the gross return at the end of one year is

$$\lim_{m\to\infty}(1+r/m)^m = e^r.$$

More generally, if the initial capital is C, invested in a bond that compounds continuously the interest at annual rate r, then the value of the investment at time t is

$$C\exp(rt).$$

Hence the log return per annum is r, which is the logarithm of the gross return. This indicates that the simple annual interest rate r quoted in the market is in fact the annual log return if the interest is compounded continuously. Note that if the interest is only paid once at the end of the year, the simple return will be r, and the log return will be $\log(1+r)$ which is always smaller than r.

In summary, a simple annual interest rate quoted in the market has two interpretations: it is the simple annual return if the interest is only paid once at the end of the year, and it is the annual log return if the interest is compounded continuously.

1.1.4 Adjustment for dividends

Many assets, for example some blue-chip stocks, pay dividends to their shareholders from time to time. A dividend is typically allocated as a fixed amount of cash per share. Therefore adjustments must then be made in computing returns to account for the contribution towards the earnings from dividends. Let D_t denote the dividend payment between time $t-1$ and t. Then the returns are now defined as follows:

$$R_t = (P_t + D_t)/P_{t-1} - 1, \quad r_t = \log(P_t + D_t) - \log P_{t-1},$$

$$R_t(k) = \big(P_t + D_t + \cdots + D_{t-k+1}\big)/P_{t-k} - 1,$$

$$r_t(k) = r_t + \cdots + r_{t-k+1} = \sum_{j=0}^{k-1} \log\left(\frac{P_{t-j} + D_{t-j}}{P_{t-j-1}}\right).$$

The above definitions are based on the assumption that all dividends are cashed out and are not re-invested in the asset.

1.1.5 Bond yields and prices

Bonds are quoted in annualized yields. A so-called zero-coupon bond is a bond bought at a price lower than its face value (also called par value or principal), with the face value repaid at the time of maturity. It does not make periodic interest payments (i.e. coupons), hence the term 'zero-coupon'. Now we consider a zero-coupon bond with the face value \$1. If the current yield is r_t and the remaining duration is D units of time, with continuous compounding, its current price B_t should satisfy the condition

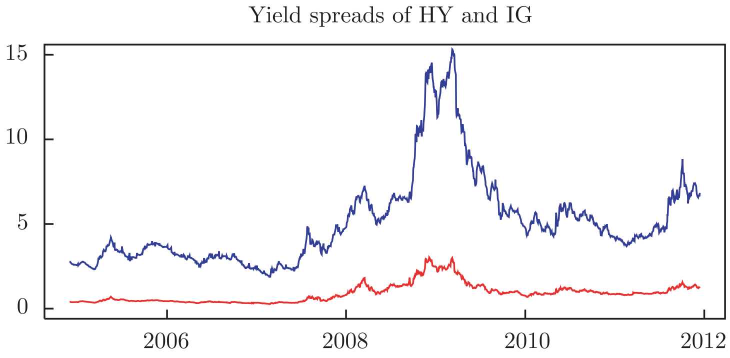
$$B_t \exp(Dr_t) = \$1,$$

i.e. the price is $B_t = \exp(-Dr_t)$ dollars. Thus, the annualized log-return of the bond is

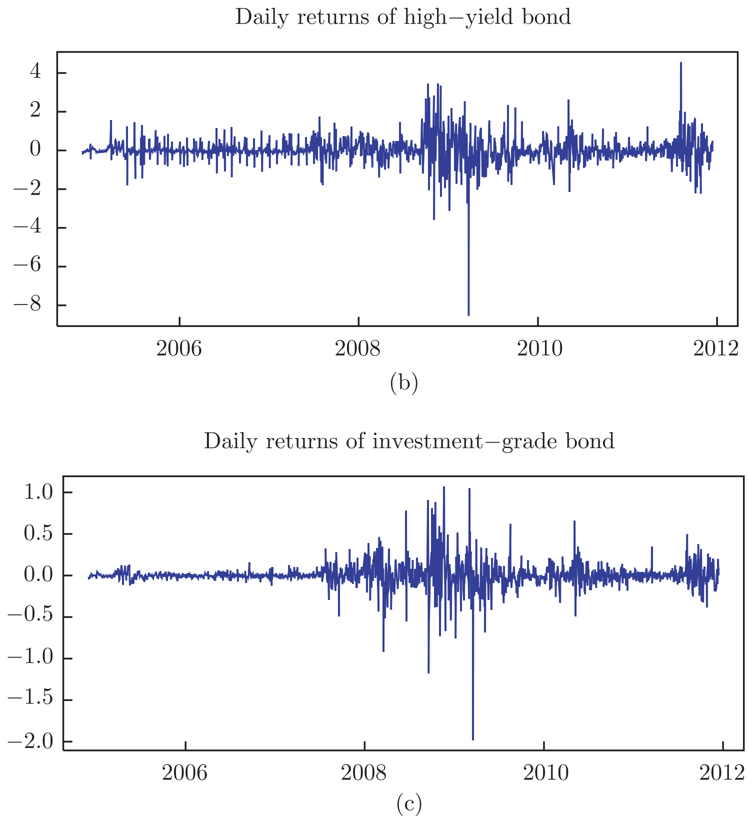
$$\log(B_{t+1}/B_t) = D(r_t - r_{t+1}). \tag{1.7}$$

Here, we ignore the fact that B_{t+1} has one unit of time shorter maturity than B_t.

Suppose that we have two baskets of high-yield bonds and investment-grade bonds (i.e. the bonds with relatively low risk of default) with an average duration of 4.4 years each. Their yields spread (i.e. the difference) over the Treasury bond with similar maturity are quoted and plotted in Figure 1.2. The daily returns of bonds can then be deduced from (1.7), which is the change of yields multiplied by the duration. The daily changes of treasury bonds are typically much smaller. Hence, the changes of yield spreads can directly be used as proxies of the changes of yields. As expected, the high-yield bonds have higher yields than the investment grade bonds, but have higher volatility too (about 3 times). The yield spreads widened significantly in a period after the financial crisis following Lehman Brothers filing bankrupt protection on September 15, 2008, reflecting higher default risks in corporate bonds.



(a)



Daily returns of high−yield bond

(b)

Daily returns of investment−grade bond

(c)

Figure 1.2 Time series of the yield spreads (the top panel) of high-yield bonds (blue curve) and investment-grade bonds (red curve), and their associated daily returns (the 2nd and 3rd panels) in November 29, 2004 – December 10, 2014.

1.1.6 Excess returns

In many applications, it is convenient to use an *excess return*, which is defined in the form $r_t - r_t^\star$, where $r_t^\star$ is a reference rate. The commonly used reference rates are, for example, bank interest rates, *LIBOR* rates (London Interbank Offered Rate: the average interest rate that leading banks in London charge when lending to other banks), log returns of a riskless asset (e.g., yields of short-term government bonds such as the 3-month US treasury bills) or market portfolio (e.g. the S&P 500 index or CRSP value-weighted index, which is the value-weighted index of all stocks traded in three major stock exchanges, created by the Center for Research in Security Prices of University of Chicago).

For bonds, *yield spread* is an excess yield defined as the difference between the yield of a bond and the yield of a reference bond such as a US treasury bill with a similar maturity.

1.2 Behavior of financial return data

In order to build useful statistical models for financial returns, we collect some empirical evidence first. To this end, we look into the daily closing indices of the S&P 500 and the daily closing share prices (in US dollars) of the Apple Inc in the period of January 1985 – February 2011. The data were adjusted for all splits and dividends, and were downloaded from *Yahoo!Finance*.

The S&P 500 is a value-weighted index of the prices of the 500 large-cap common stocks actively traded in the United States. Its present form has been published since 1957, but its history dates back to 1923 when it was a value-weighted index based on 90 stocks. It is regarded as a bellwether for the American economy. Many mutual funds, exchange-traded funds, pension funds etc are designed to track the performance of S&P 500. The first panel in Figure 1.3 is the time series plot for the daily closing indices of S&P 500. It shows clearly that there was a slow and steady increase momentum in 1985-1987. The index then reached an all-time high on March 24, 2000 during the dot-com bubble, and consequently lost about 50% of its value during the stock market downturn in 2002. It peaked again on October 9, 2007 before suffering during the credit crisis stemming from subprime mortgage lending in 2008-2010. The other three panels in Figure 1.3 show the daily, the weekly and the monthly log returns of the index. Although the profiles of the three plots are similar, the monthly return curve is a 'smoothed' version of, for example, the daily return curve which exhibits higher volatile fluctuations. In particular, the high volatilities during the 2008-2010 are more vividly depicted in the daily return plot. In contrast to the prices, the returns oscillate around a constant level close to 0. Furthermore, high oscillations tend to cluster together, reflecting more volatile market periods. Those features on return data are also apparent in the Apple stock displayed in Figure 1.4. The share prices of the Apple Inc are also non-stationary in time in the sense that the price movements over different time periods are different. For example in 1985 – 1998, the prices almost stayed at a low level. Then it experienced a steady increase until September 29, 2000 when the Apple's value sliced in half due to the earning warning in the last quarter of the year. The more recent surge of the price increase was largely due to Apple's success in the mobile consumer electronics market via its products the iPod, iPhone and iPad, in spite of its fluctuations during the subprime mortgage credit crisis.

We plot the normalized histograms of the daily, the weekly and the monthly log returns of the S&P 500 index in Figure 1.5. For each histogram, we also superimpose the normal density function with the same mean and variance. Also plotted in Figure 1.5 are the quantile-quantile plots for the three returns. (See an introduction on Q-Q plots in Section 1.5.) It is clear that the returns within the given holding

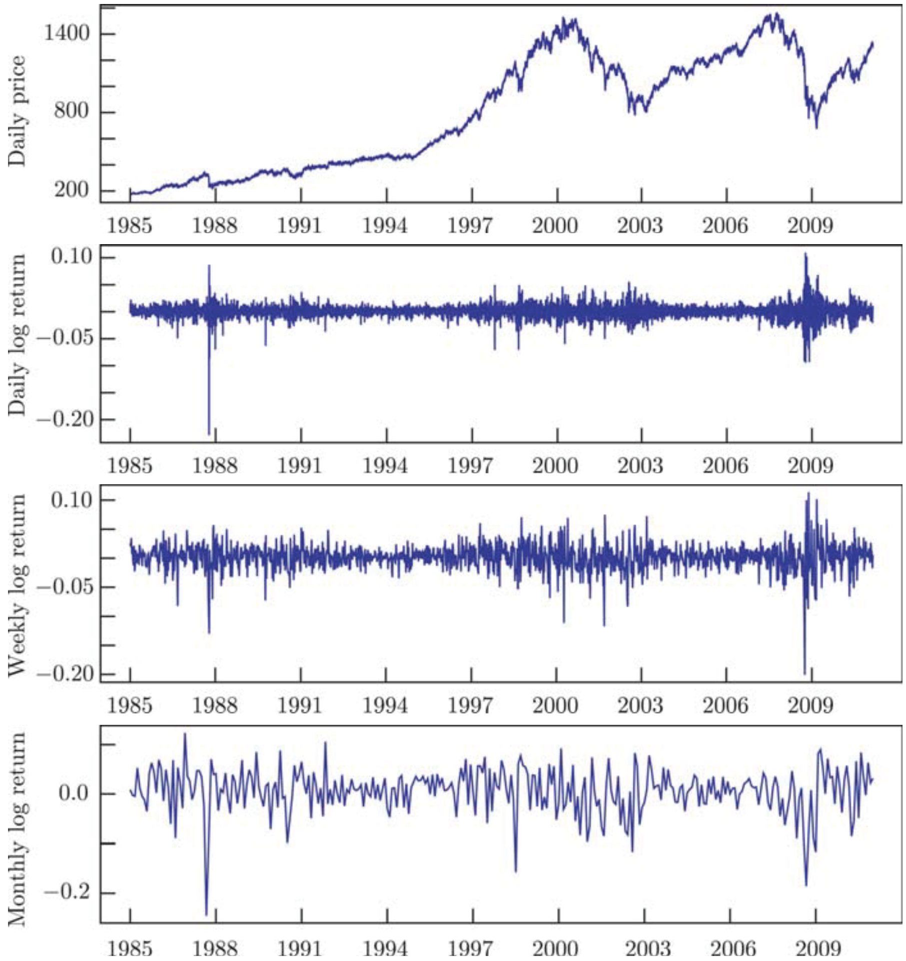
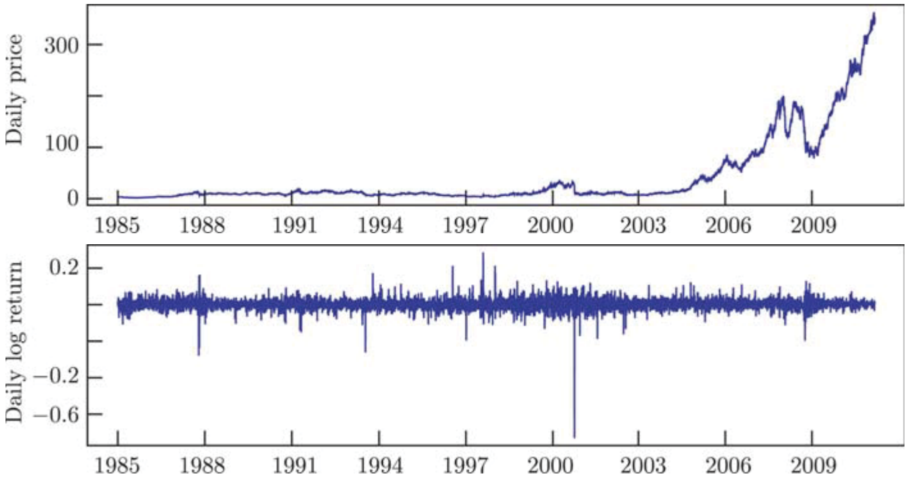


Figure 1.3 Time series plots of the daily indices, the daily log returns, the weekly log returns, and the monthly log returns of S&P 500 index in January 1985 – February 2011.



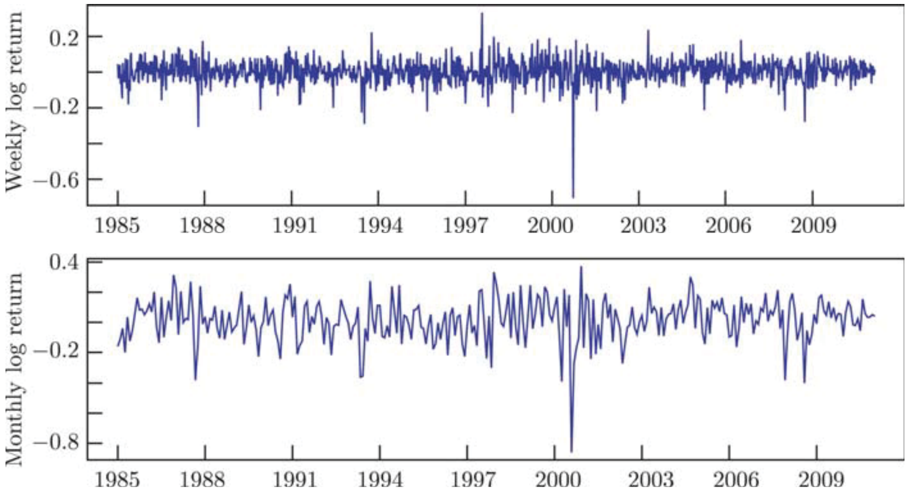


Figure 1.4 Time series plots of the daily prices, the daily log returns, the weekly log returns, and the monthly log returns of the Apple stock in January 1985 – February 2011.

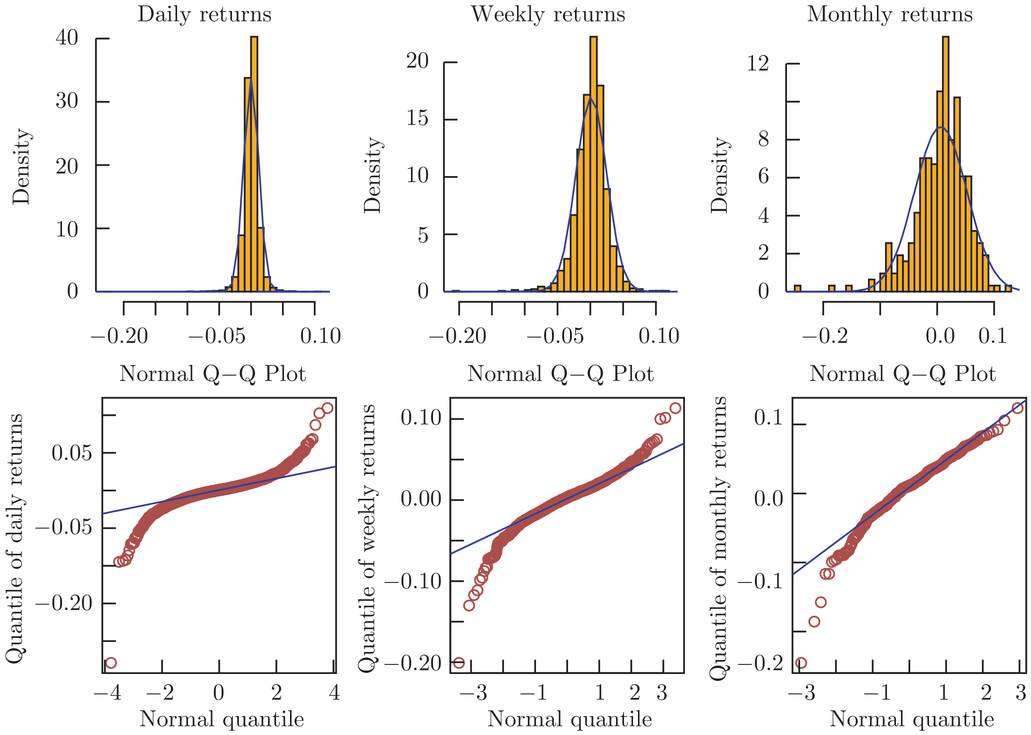


Figure 1.5 Histograms (the top panels) and Q-Q plots (the bottom panels) of the daily, weekly, and monthly log returns of S&P 500 in January 1985 – February 2011. The normal density with the same mean and variance are superimposed on the histogram plots.

periods are not normally distributed. Especially the tails of the return distributions are heavier than those of the normal distribution, which is highlighted explicitly in the Q-Q plots: the left tail (red circles) is below (negatively larger) the blue line, and the right tail (red circles) is above (larger) the blue line. We have also noticed that when the holding period increases from a day, a week to a month, the tails of the distributions become lighter. In particular the upper tail of the distribution for the monthly returns is about equally heavy as that of a normal distribution (red circles and blue line are about the same). All the distributions are skewed to the left due to a few large negative returns. The histograms also show that the distribution for the monthly returns is closer to a normal distribution than those for the weekly returns and the daily returns. The similar patterns are also observed in the Apple return data; see Figure 1.6.

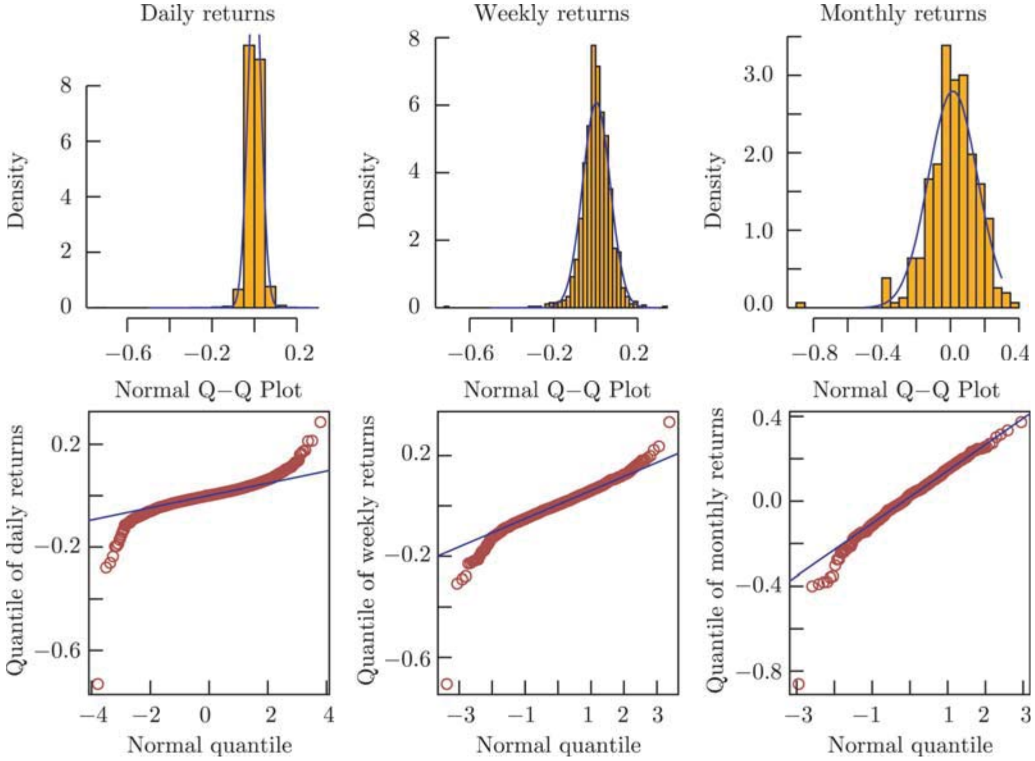


Figure 1.6 Histograms (the top panels) and Q-Q plots (the bottom panels) of the daily, weekly, and monthly log returns of the Apple stock in January 1985 – February 2011. The normal density with the same mean and variance are superimposed on the histogram plots.

Figures 1.7 and 1.8 plot the sample autocorrelation function (ACF) $\widehat{\rho}_k$ against the time lag k for the log returns, the squared log returns and the absolute log returns. Given a return series $r_1, \cdots, r_T$, the sample autocorrelation function is defined as $\widehat{\rho}_k = \widehat{\gamma}_k / \widehat{\gamma}_0$, where

$$\widehat{\gamma}_k = \frac{1}{T}\sum_{t=1}^{T-k}(r_t - \bar{r})(r_{t+k} - \bar{r}), \quad \bar{r} = \frac{1}{T}\sum_{t=1}^{T} r_t. \tag{1.8}$$

$\widehat{\gamma}_k$ is the sample autocovariance at lag k. It is (about) the same as the sample correlation coefficient of the paired observations $\{(r_t, r_{t+k})\}_{t=1}^{T-k}$ (the difference is in the definition of the sample mean in the calculation of the sample covariance). The sample autocorrelation functions for the squared and the absolute returns are defined in the same manner but with r_t replaced by, respectively, r_t^2 and $|r_t|$. For each ACF plot in Figures 1.7 and 1.8, the two dashed horizontal lines, which are $\pm 1.96/\sqrt{T}$, are the bounds for the 95% confidence interval for ρ_k if the true value is $\rho_k = 0$. Hence ρ_k would be viewed as not significantly different from 0 if its estimator $\widehat{\rho}_k$ is between those two lines. It is clear from Figures 1.7 and 1.8 that all the daily, weekly and monthly returns for both S&P 500 and the Apple stock exhibit no significant autocorrelation, supporting the hypothesis that the returns of a financial asset are uncorrelated across time. However there are some small but significant autocorrelations in the squared returns and more in the absolute returns.

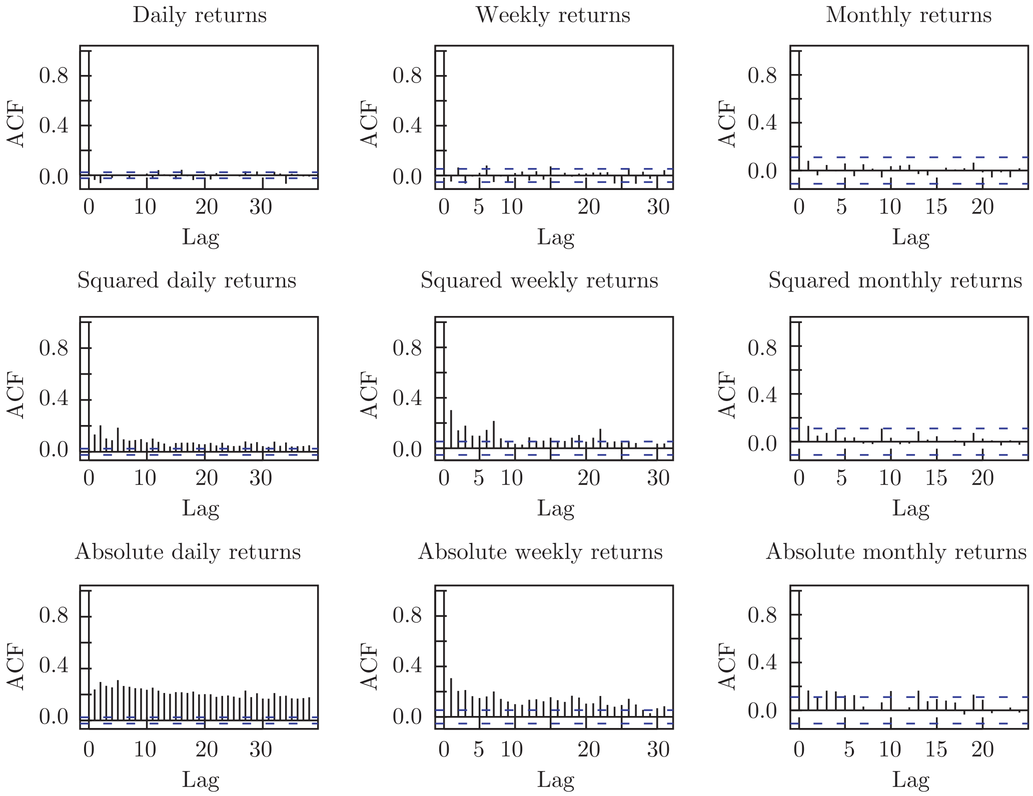


Figure 1.7 Autocorrelations of the daily, weekly, and monthly log returns, the squared daily, weekly, and monthly log returns, and the absolute daily, weekly, and monthly log returns of S&P 500 in January 1985 – February 2011.

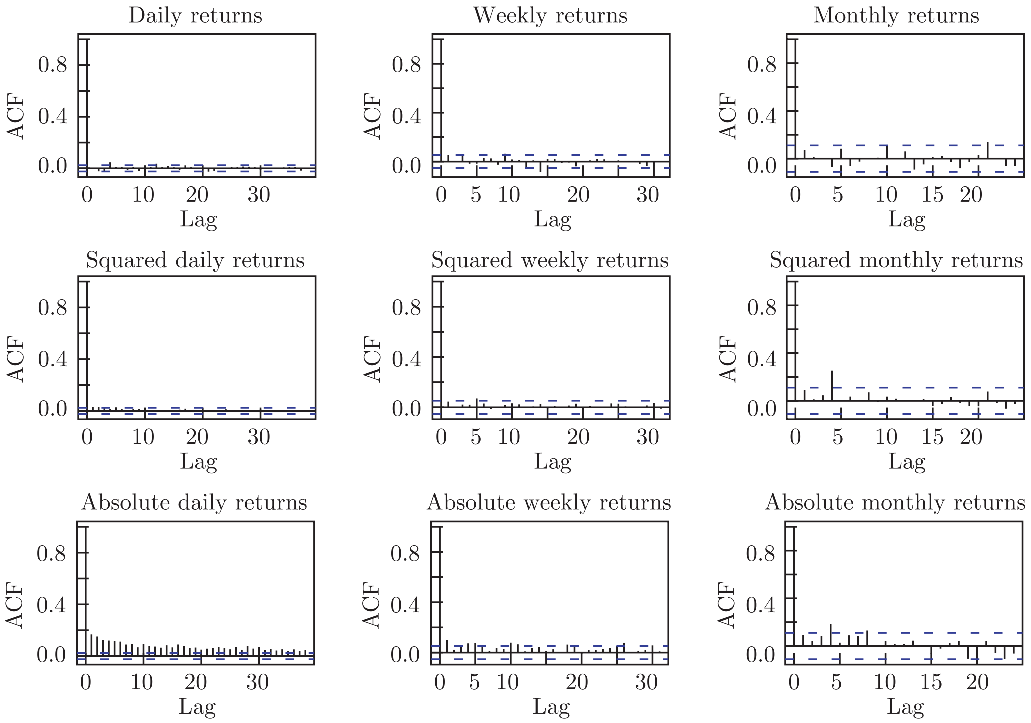


Figure 1.8 Autocorrelations of the daily, weekly, and monthly log returns, the squared daily, weekly, and monthly log returns, and the absolute daily, weekly, and monthly log returns of the Apple stock in January 1985 – February 2011.

Furthermore the autocorrelations are more pronounced and more persistent in the daily data than in weekly and monthly data. Since the correlation coefficient is a measure of linear dependence, the above empirical evidence indicates that the returns of a financial asset are linearly independent with each other, although there exist nonlinear dependencies among the returns at different lags. Especially the daily absolute returns exhibit significant and persistent autocorrelations — a characteristic of so-called long memory processes.

1.2.1 Stylized features of financial returns

The above findings from the two real data sets are in line with the so-called stylized features in financial returns series, which are observed across different kinds of assets including stocks, portfolios, bonds and currencies. See, e.g. Rydberg (2000). We summarize these features below.

(i) *Stationarity.* The prices of an asset recorded over times are often not stationary due to, for example, the steady expansion of economy, the increase of productivity resulting from technology innovation, and economic recessions or financial crisis. However their returns, denoted by r_t for $t \geqslant 1$, typically fluctuates around a constant level, suggesting a constant mean over time. See Figures 1.3 and 1.4. In fact most return sequences can be modeled as a stochastic processes with at least time-invariant first two moments (i.e. the weak stationarity; see 2.1). A simple (and perhaps over-simplistic) approach is to assume that all the finite dimensional distributions of a return sequence are time-invariant.

(ii) *Heavy tails.* The probability distribution of return r_t often exhibits heavier tails than those of a normal distribution. Figures 1.5 and 1.6 provide the *quantile-quantile plot* or *Q-Q plot* for graphical checking of normality. See Section 1.5 for detail. A frequently used statistic for checking the normality (including tail-heaviness) is the Jarque-Bera test presented in Section 1.5. Nevertheless r_t is assumed typically to have at least two finite moments (i.e. $E(r_t^2) < \infty$), although it is debatable how many moments actually exist for a given asset.

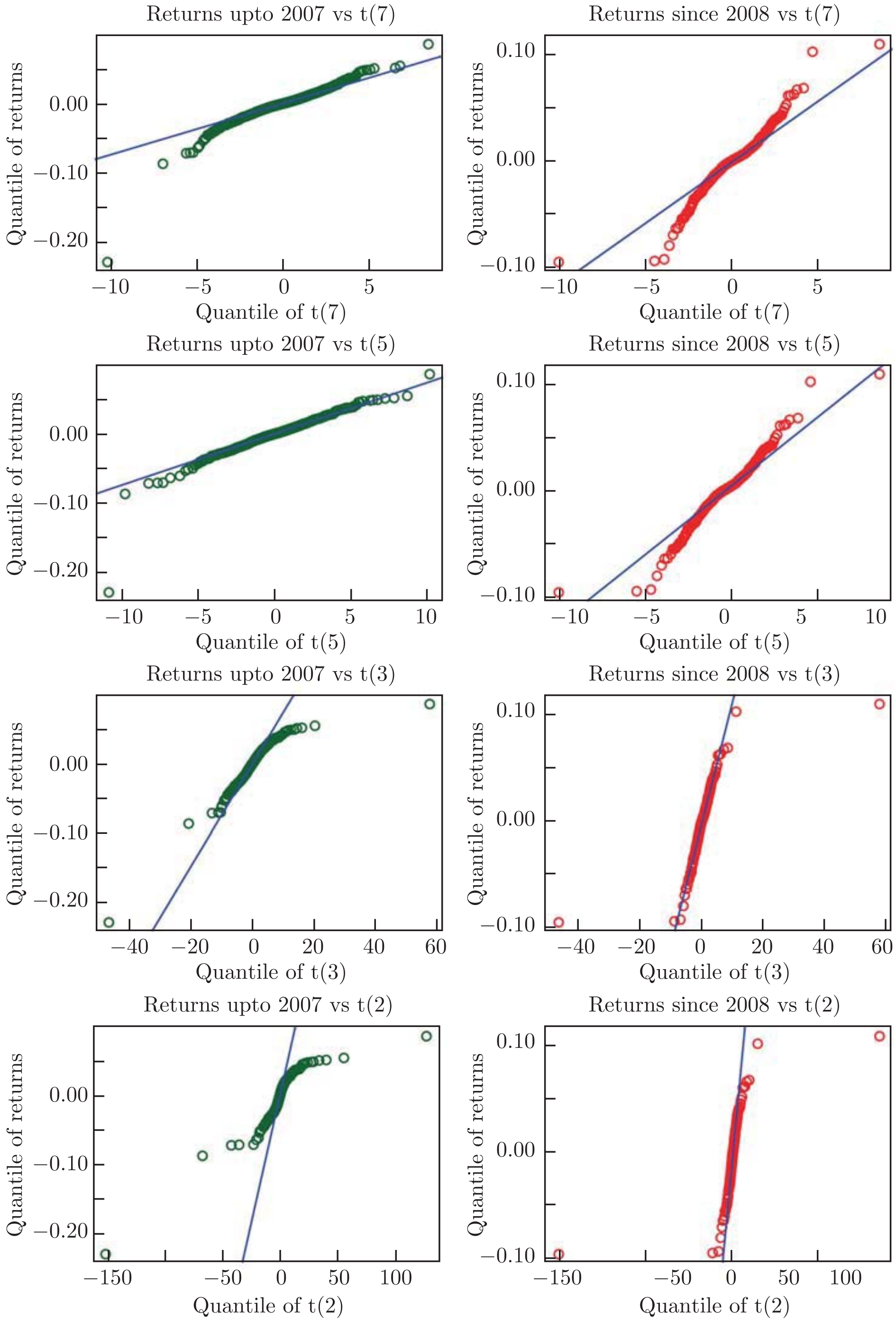
The density of *t-distribution* with degree of freedom ν is given by

$$f_\nu(x) = d_\nu^{-1} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}, \tag{1.9}$$

where $d_\nu = B(0.5, 0.5\nu)\sqrt{\nu}$ is the normalization constant and B is a beta function. This distribution is often denoted as $t(\nu)$ or t_ν. Its tails are of polynomial order $f_\nu(x) \asymp |x|^{-(\nu+1)}$ (as $|x| \to \infty$), which are heavier than the normal density. Note that for any random variable $X \sim t(\nu)$, $E\{|X|^\nu\} = \infty$ and $E\{|X|^{\nu-\delta}\} < \infty$ for any $\delta \in (0, \nu]$.

When ν is large, $t(v)$ is close to a normal distribution. In fact, based on a sample of size 2500 (approximately 10-year daily data), one can not differentiate $t(10)$ from a normal distribution based on, for example, the Kolmogorov-Smirnov test (the function `KS.test` in R). However their tail behaviors are very different: A 5-standard-deviation (SD) event occurs once in every 14000 years under a normal distribution, once in every 15 years under t_{10}, and once in every 1.5 years under $t_{4.5}$. The calculation goes as follows. The probability of getting a -5 SD daily shock or worse under the normal distribution is 2.8665×10^{-7} (which is $P(Z < -5)$ for $Z \sim N(0,1)$), or 1 in 3488575 days. Dividing this by approximately 252 trading days per year yields the result of 13844 years. A similar calculation can be done with different kinds of t-distributions. If the tails of stock returns behave like $t_{4.5}$, the left tail of typical daily S&P 500 returns, the occurrence of -5 SD event is more often than what we would conceive.

Figure 1.9 plots the quantiles of the S&P 500 returns in the period January 1985 – February 2011 against the quantiles of $t(\nu)$ distributions with $\nu = 2, 3, \cdots, 7$. It is clear that the tails of the S&P 500 return are heavier than the tails of $t(5)$, and are thinner than the tails of $t(2)$ (and perhaps also $t(3)$). Hence it is reasonable to assume that the second moment of the return of the S&P 500 is finite while the 5th moment should be infinity.


Returns upto 2007 vs t(7)

Returns since 2008 vs t(7)

Returns upto 2007 vs t(5)

Returns since 2008 vs t(5)

Returns upto 2007 vs t(3)

Returns since 2008 vs t(3)

Returns upto 2007 vs t(2)

Returns since 2008 vs t(2)

Figure 1.9 Plots of log returns against simple returns of the Apple Inc share prices in January 1985 – February 2011. The blue straight lines mark the positions where the two returns are identical.

(iii) *Asymmetry.* The distribution of return r_t is often negatively skewed (Figures 1.5 and 1.6), reflecting the fact that the downturns of financial markets are often much steeper than the recoveries. Investors tend to react more strongly to negative news than to positive news.

(iv) *Volatility clustering.* This term refers to the fact that large price changes (i.e. returns with large absolute values) occur in clusters. See Figures 1.3 and 1.4. Indeed, large price changes tend to be followed by large price changes, and periods of tranquility alternate with periods of high volatility. The time varying volatility can easily be see in Figures 1.3. The standard deviation of S&P 500 returns from November 29, 2004 to December 31, 2007 (3 months before JP Morgan Chase offered to acquire Bear Stearns at a price of $2 per share on March 17, 2008) is 0.78%, and the standard deviation of the returns since 2008 is 1.83%. The volatility of S&P 500 in 2005 and 2006 is merely 0.64%, whereas the volatility at the height of the 2008 financial crisis (September 15, 2008 to March 16, 2009, i.e. from a month before to 6 months after Lehmann Brother's fall) is 3.44%.

(v) *Aggregational Gaussianity.* Note that a return over k days is simply the aggregation of k daily returns; see (1.6). When the time horizon increases, the central limit law sets in and the distribution of the returns over a long time-horizon (such as a month) tends toward a normal distribution. See Figures 1.5 and 1.6.

(vi) *Long range dependence.* The returns themselves hardly show any serial correlation, which, however, does not mean that they are independent. In fact, both daily squared and absolute returns often exhibit small and significant autocorrelations. Those autocorrelations are persistent for absolute returns, indicating possible long-memory properties. It is also noticeable that those autocorrelations become weaker and less persistent when the sampling interval is increased from a day, to a week to a month. See Figure 1.7.

(vii) *Leverage effect.* Asset returns are negatively correlated with the changes of their volatilities (Black 1976, Christie 1982). As asset prices decline, companies become more leveraged (debt to equity ratios increase) and riskier, and hence their stock prices become more volatile. On the other hand, when stock prices become more volatile, investors demand high returns and hence stock prices go down. Volatilities caused by price decline are typically larger than the appreciations due to declined volatilities. To examine the leverage effect, we use the VIX, which is a proxy of the implied volatility of a basket of at-money S&P 500 options maturity in one month, as the proxy of volatility of the S&P 500 index. Figure 1.10 shows the time series of VIX and S&P 500 index (left panel) and the returns of S&P 500 index against the change of volatilities (right panel). The leverage effect is strong, albeit VIX is not a prefect measure of the volatility of the S&P 500 index, involving the volatility risk premium (Ait-Sahalia, Fan, and Li, 2013)

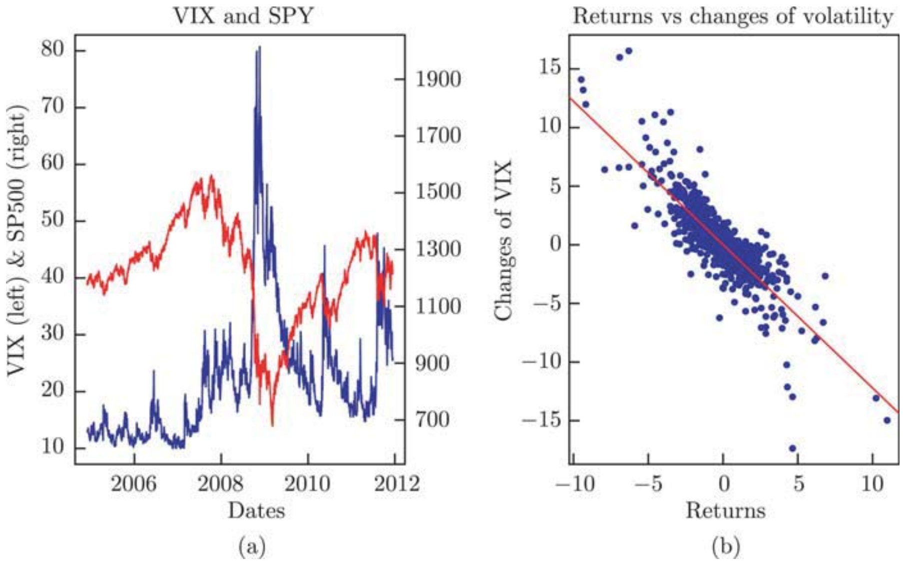


Figure 1.10 Time series plot of VIX (red) and the S&P 500 index (blue) in Nov. 29, 2004 – Dec. 14, 2011 (the left panel), and the plot of the daily S&P 500 returns (in percent) against the changes of VIX (the right panel).

1.3 Efficient markets hypothesis and statistical models for returns

The efficient markets hypothesis (EMH) in finance assumes that asset prices are fair, information is accessible for everybody and is assimilated rapidly to adjust prices, and people (including traders) are rational. Therefore price P_t incorporates all relevant information up to time t, individuals do not have comparative advantages in the acquisition of information. The price change $P_t - P_{t-1}$ is *only* due to the arrival of "news" between t and $t+1$. Hence individuals have no opportunities for making an investment with return greater than a fair payment for undertaking riskiness of the asset. A shorthand for the EHM: *the price is right*, and *there exist no arbitrage opportunities*.

The above describes the strong form of the EMH: security prices of traded assets reflect instantly all available information, public or private. A semi-strong form states that security prices reflect efficiently all public information, leaving rooms for the value of private information. The weak form merely assumes security prices reflect all past publicly available information.

Under the EHM, an asset return process may be expressed as

$$r_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim (0, \sigma_t^2), \tag{1.10}$$

where μ_t is the *rational expectation* of r_t at time $t-1$, and ε_t represents the return due to unpredictable "news" which arrives between time $t-1$ and t. In this sense, ε_t is an *innovation* – a term used very often in time series literature. We use the notation $X \sim (\mu, \nu)$ to denote that a random variable X has mean μ and variance ν, respectively. The assumption that $E\varepsilon_t = 0$ reflects the belief that on average the actual change of log price equals the expectation μ_t.

By combining the EHM model (1.10) with some stylized features outlined in Section 1.2, most frequently used statistical models for financial returns admit the form

$$r_t = \mu + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{WN}(0, \sigma^2), \tag{1.11}$$

where $\mu = Er_t$ is the expected return, which is assumed to be a constant. The notation $\varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$ denotes that $\varepsilon_1, \varepsilon_2, \cdots$ form a white noise process with $E\varepsilon_t = 0$ and $\mathrm{var}(\varepsilon_t) = \sigma^2$; see (i) below. Here we assume that $\mathrm{var}(\varepsilon_t) = \mathrm{var}(r_t) = \sigma^2$ is a finite positive constant, noting that *most* varying volatilities in the return plots in Figures 1.3 & 1.4 can be represented by conditional heteroscadasticity under the martingale difference assumption (ii) below. The assumption (1.11) is reasonable, supported by the empirical analysis in section 1.2.

There are three different types of assumptions about the innovations $\{\varepsilon_t\}$ in (1.11), from the weakest to the strongest.

(i) *White noise innovations*: $\{\varepsilon_t\}$ are white noise, denoted as $\varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$. Under this assumption, $\mathrm{Corr}(\varepsilon_t, \varepsilon_s) = 0$ for all $t \neq s$.

(ii) *Martingale difference innovations*: ε_t form a martingale difference sequence in the sense that for any t

$$E(\varepsilon_t | r_{t-1}, r_{t-2}, \cdots) = E(\varepsilon_t | \varepsilon_{t-1}, \varepsilon_{t-2}, \cdots) = 0. \tag{1.12}$$

One of the most frequently used format for martingale difference innovations is of the form

$$\varepsilon_t = \sigma_t \eta_t, \tag{1.13}$$

where $\eta_t \sim \mathrm{IID}(0, 1)$ (see (iii) below), and σ_t is a predictable volatility process, known at time $t-1$, satisfying the condition

$$E(\sigma_t | r_{t-1}, r_{t-2}, \cdots) = \sigma_t.$$

Note that ARCH and GARCH processes are special cases of (1.13).

(iii) *IID innovations*: ε_t are independent and identically distributed, denoted as $\varepsilon_t \sim \mathrm{IID}(0, \sigma^2)$.

The assumption of IID innovations is the strongest. It implies that the innovations are martingale differences. On the other hand, if $\{\varepsilon_t\}$ satisfies (1.12), it holds that for any $t > s$,

$$\begin{aligned}\operatorname{cov}(\varepsilon_t, \varepsilon_s) &= E(\varepsilon_t \varepsilon_s) = E\{E(\varepsilon_t \varepsilon_s | \varepsilon_{t-1}, \varepsilon_{t-2}, \cdots)\} \\ &= E\{\varepsilon_s E(\varepsilon_t | \varepsilon_{t-1}, \varepsilon_{t-2}, \cdots)\} = 0.\end{aligned}$$

Hence, $\{\varepsilon_t\}$ is a white noise series. Therefore the relationship among the three types of innovations is as follows:

$$\text{IID} \Rightarrow \text{Martingale differences} \Rightarrow \text{White noise.}$$

The relationship is summarized in Figure 1.11.

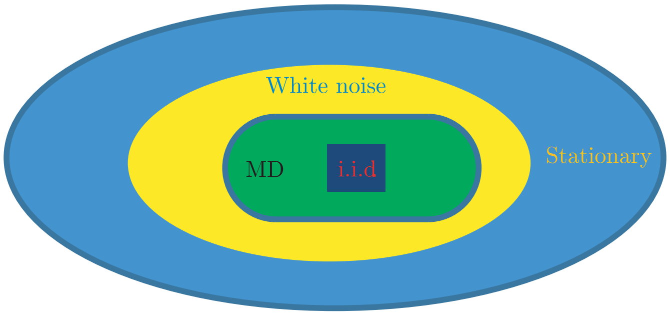


Figure 1.11 Relationship among different processes: Stationary processes are the largest set, followed by white noise, martingale difference (MD), and i.i.d. processes. There are many useful processes between stationary processes and white noise processes, to be detailed in Chapters 2 and 3.

The white noise assumption is widely observed in financial return data. It is consistent with the stylized features presented in Figures 1.7 and 1.8. It is implied by the EMH, as the existence of the non-zero correlation between ε_{t+1} and its lagged values leads to an improvement on the prediction for r_{t+1} over the rational expectation μ. This violates the hypothesis that ε_{t+1} is unpredictable at time t. To illustrate this point, suppose $\operatorname{Corr}(\varepsilon_{t+1}, \varepsilon_s) = \rho \neq 0$ for some $s \leqslant t$. Then $\widetilde{r}_{t+1} = \mu + \rho(r_s - \mu)$ is a legitimate predictor for r_{t+1} at time t. Under the EMH, the fair predictor for r_{t+1} at time t is $\widehat{r}_{t+1} = \mu$. Then it is easy to see that

$$E\{(\widehat{r}_{t+1} - r_{t+1})^2\} = \operatorname{var}(\varepsilon_{t+1}) = \sigma^2,$$

whereas

$$E\{(\widetilde{r}_{t+1} - r_{t+1})^2\} = E\{(\rho\varepsilon_s - \varepsilon_{t+1})^2\} = (1 - \rho^2)\sigma^2 < \sigma^2.$$

i.e. the mean squared predictive error of $\widetilde{r}_{t+1}$ is smaller. Hence the white noise assumption is appropriate and arguably necessary under the EMH. It states merely that the asset returns cannot be predicted by any linear rules. However it says nothing beyond the first two moments and remains silent on the question whether the traded asset returns can be predicted by nonlinear rules or by other complicated strategies.

On the other hand, the empirical evidence reported in Section 1.2 indicates that the IID assumption is too strong and too restrictive to be true in general. For example, the squared and the absolute returns of both S&P 500 index and the Apply stock exhibit significant serial correlations, indicating that $r_1, r_2, \cdots$, therefore also $\varepsilon_1, \varepsilon_2, \cdots$, are not independent with each other; see Figures 1.7 & 1.8.

Note that $r_t = \log(P_t/P_{t-1})$. It follows from (1.11) that

$$\log P_t = \mu + \log P_{t-1} + \varepsilon_t. \tag{1.14}$$

Hence under the assumption that the innovations ε_t are IID, the log prices $\log P_t,\ t = 1, 2, \cdots$ form a *random walk*, and the prices $P_t,\ t = 0, 1, 2, \cdots$, are a geometric random walk. Since the future is independent of the present and the past, the EMH holds in the most strict sense and nothing in the future can be predicted based on the available information up to the present. If we further assume ε_t to be normal, P_t follows a log normal distribution. Then the price process P_t, $t = 0, 1, 2, \cdots$, is a log normal geometric random walk. As the length of time unit shrinks to zero, the number of periods goes to infinity and the appropriately normalized random walk $\log P_t$ converges to a Brownian motion, and the geometric random walk P_t converges to a geometric Brownian motion under which the celebrated Black-Scholes formula is derived. The concept that stock market prices evolve according to a random walk can be traced back at least to French mathematician Louis Bachelier in his PhD dissertation in 1900.

A weaker form of random walks relaxes ε_t to be, for example, martingale differences. The martingale difference assumption offers a middle ground between white noise and IID. While retaining the white noise (i.e. the linear independence) property, it does not rule out the possibility of some nonlinear dependence, i.e. $\{r_t\}$ are uncorrelated but $\{r_t^2\}$ or $\{|r_t|\}$ may be dependent with each other. Under this assumption model (1.11) may accommodate conditional heteroscadasticity as in (1.13). In fact, many volatility models including ARCH, GARCH and stochastic volatility models are special cases of (1.11) and (1.13) with ε_t being martingale differences.

The martingale difference assumption retains the hypothesis that the innovation ε_{t+1} is unpredictable at time t at least as far as the point prediction is concerned. (Later we will learn that the interval predictions for ε_{t+1}, or more precisely, the risks ε_{t+1} may be better predicted by incorporating the information from its lagged values.) The best point predictor for r_{t+1} based on $r_t, r_{t-1}, \cdots$ is the conditional expectation

$$\widehat{r}_{t+1} = E(r_{t+1}|r_t, r_{t-1}, \cdots) = \mu + E(\varepsilon_{t+1}|\varepsilon_t, \varepsilon_{t-1}, \cdots) = \mu,$$

which is the fair expectation of r_{t+1} under the EMH. The last equality in the above expression is guaranteed by (1.12). We call $\widehat{r}_{t+1}$ as the best in the sense that it minimizes the mean squared predictive errors among all the point predictors based on $r_t, r_{t-1}, \cdots$. See Section 2.9.1 for additional details.

In summary, the *martingale hypothesis*, which postulates model (1.11) with martingale difference sequence $\{\varepsilon_t\}$, assures that the returns of assets cannot be predicted by any rules, but allow volatility to be predictable. It is the most appropriate mathematical form of the efficient market hypothesis.

1.4 Tests related to efficient markets hypothesis

One fundamental question in financial econometrics is if the efficient markets hypothesis is consistent with empirical data. One way to verify this hypothesis is to test if returns are predictable. In the sequel we will introduce two statistical tests which address this issue from different angles.

1.4.1 Tests for white noise

From the discussion in the previous section, we have learned that if returns are unpredictable, they should be at least white noise. On the other hand, the assumption that returns are IID is obviously too strong. The autocorrelations in squared and absolute returns shown in Figures 1.7 & 1.8 clearly indicate that returns at different times are not independent of each other. In spite of a large body of statistical tests for IID (e.g. see the rank-based test of Hallin and Puri (1988), and also see section 2.2 of Campbell et al. (1997), we focus on testing white noise hypothesis, i.e. returns are linearly independent but may depend on each other in some nonlinear manners. The test for white noise is one of the oldest and the most important tests in statistics, as many testing problems in linear modelling may be transformed into a white noise test. There exist quite a few testing methods; see section 7.4 of Fan and Yao (2003) and the references within. We introduce below a simple and frequently used omnibus test, i.e. *Ljung-Box portmanteau test*.

The linear dependence between r_t and r_{t-k} is comprehensively depicted by the correlation function between r_t and r_{t-k}:

$$\rho_k \equiv \text{Corr}(r_t, r_{t-k}) = \frac{\text{cov}(r_t, r_{t-k})}{\sqrt{\text{var}(r_t)\text{var}(r_{t-k})}}.$$

In fact if $\rho_k = 0$, r_t and r_{t-k} are linearly independent, and $\rho_k = \pm 1$ if and only if $r_t = a + br_{t-k}$ for some constants a and b. When $\{r_t\}$ is a white noise sequence, $\rho_k = 0$ for all $k \neq 0$. In practice, we do not know ρ_k. Based on observed returns $r_1, \cdots, r_T$, we use the estimator $\widehat{\rho}_k = \widehat{\gamma}_k / \widehat{\gamma}_0$ instead, where $\widehat{\gamma}_k$ is defined in (1.8). The Ljung-Box Q_m-statistic is defined as

$$Q_m = T(T+2) \sum_{j=1}^{m} \frac{1}{T-j} \widehat{\rho}_j^2, \tag{1.15}$$

where $m \geqslant 1$ is a prescribed integer. Note that Q_m is essentially a weighted sum of the squared sample ACF over the first m lags, though the weights are approximately the same when $T \gg m$. Intuitively we reject the white noise hypothesis for large values of Q_m. How large is large depends on the theoretical distribution of Q_m under the null hypothesis, which turns out to be problematic; see below. In practice a chi-square approximation is used: For $\alpha \in (0, 1)$, let $\chi^2_{\alpha,m}$ denote the top α-th percentile of the χ^2-distribution with m degrees of freedom.

The Ljung-Box portmanteau test: Reject the hypothesis that $\{r_t\}$ is a white noise at the significance level α if $Q_m > \chi^2_{\alpha,m}$ or its P-value, computed as $P(Q > Q_m)$ with $Q \sim \chi^2_m$, is smaller than α.

In practice one needs to choose m in (1.15). Conventional wisdom suggests to use small m, as serial correlation is often at its strongest at small lags. This also avoids the large estimation errors in sample ACF at large lags and error accumulation issue in summation (1.15). However using too small m may miss the autocorrelations beyond the lag m. It is not uncommon in practice that the Ljung-Box test is performed simultaneously with different values of m.

The *R*-function to perform the Ljung-Box test is `Box.test(x, lag=m, type="Ljung")`, where `x` is a data vector. The command `Box.test(x, lag=m, type="Box")` performs the Box-Pierce test with the statistic Q^*_m.

We apply the Ljung-Box test to the monthly log returns of the S&P 500 index displayed in the bottom panel of Figure 1.3. The sample size is $T = 313$. The testing results are shown in Table 1.1. We cannot reject the white noise hypothesis for the log return data, as the tests with $m = 1, 6, 12$ and 24 are all not significant. In contrast, the tests for the absolute returns are statistically significant with the P-values not greater than 0.3%, indicating that the absolute returns are not white noise. The tests for the squared returns are less clearly cut with the smallest P-value 0.019 for $m = 1$ and the largest P-value 0.492 for $m = 24$. Indeed, the monthly data exhibits much weaker correlation than daily. See also the three panels on the right in Figure 1.7.

Table 1.1 P-values based on the Ljung-Box test for the S&P 500 data

	m	1	6	12	24
returns	Q_m	2.101	5.149	8.958	14.080
	P-value	0.147	0.525	0.707	0.945
squared returns	Q_m	5.517	12.292	16.964	23.474
	P-value	0.019	0.056	0.151	0.492
absolute returns	Q_m	8.687	39.283	49.721	76.446
	P-value	0.003	0.000	0.000	0.000

A different but related approach is to consider the normalized Q_m^*-statistic:

$$\frac{1}{\sqrt{2m}}\Big\{T\sum_{j=1}^{m}\widehat{\rho}(j)^2 - m\Big\}.$$

The asymptotic normality of this statistic under the condition that $m \to \infty$ and $m/T \to 0$ has been established for IID data by Hong (1996), for martingale differences by Hong and Lee (2003) (see also Durlauf (1991) and Deo (2000)), and for other non-IID white noise processes by Shao (2011), and Xiao and Wu (2011). However, those convergences are typically slow or very slow, resulting in the size distortation of the tests based on the asymptotic normality. In addition, as pointed out above, when j is large, ρ_j tends to be small. Therefore, including those terms $\hat{\rho}_j^2$ adds noises to the test statistic without increasing signals. How to choose a relevant m adds a further complication in using this approach. Horowitz et al.(2006) proposed a double blockwise bootstrap method to perform the tests with the statistic Q_m^* for non-IID white noise.

1.4.2 Remarks on the Ljung-Box test*

The chi-square approximation for the null distribution of Ljung-Box test statistic is based on the fact that when $\{r_t\}$ is an IID sequence, $\widehat{\rho}_1, \cdots, \widehat{\rho}_m$ are asymptotically independent, and each of them has an asymptotic distribution $N(0, 1/T)$ (Theorem 2.8(iii) of Fan and Yao (2003)). Hence

$$Q_m^* \equiv T\sum_{j=1}^{m}\widehat{\rho}_j^2 \sim \chi_m^2 \quad \text{approximately for large } T.$$

Now, it is easy to see that Q_m is approximately the same as Q_m^* when T is large, since $(T+2)/(T-j) \approx 1$. Hence, it also follows χ_m^2-distribution under the null hypothesis. In fact Q_m^* is the test statistic proposed by Box and Pierce (1970). However Ljung and Box (1978) subsequently discovered that the χ^2-approximation to the distribution of Q_m^* is not always adequate even for T as large as 100. They suggest to use the statistic Q_m instead as its distribution is closer to χ_m^2. See also Davies, Triggs and Newbold (1977).

A more fundamental problem in applying the Ljung-Box test is that the statistic itself is defined to detecting the departure from white noise, but the asymptotic χ^2-distribution can only be justified under the IID assumption. Therefore, as formulated above, it should not be used to test the hypothesis that the returns are white noise but not IID, as then the asymptotic null distributions of $\widehat{\rho}(k)$ depend on the high moments of the underlying distribution of r_t. These asymptotic null distributions may typically be too complicated to be directly useful in the sense that the asymptotic null distributions of Q_m or Q_m^* may then not be of the known forms for fixed m; see, e.g. Romano and Thombs (1996). Unfortunately this problem also applies to most (if not all) other omnibus white noise tests.

One alternative is to impose an explicit assumption on the structure of white noise process (such as a GARCH structure), then some resampling methods may be employed to simulate the null distribution of Q_m. Furthermore if one is also willing to impose some assumptions on the parametric form of a possible departure from white noise, a likelihood ratio test can be employed, which is often more powerful than a omnibus nonparametric test, as the latter tries to detect the departure (from white noise) to all different possibilities. The analogy is that an all-purpose tool is typically less powerful on a particular task than a customized tool. An example of the customized tool is the Dicky-Fuller test in the next section. However it itself is a challenge to find relevant assumptions. This is why omnibus tests such as the Ljung-Box test are often used in practice, in spite of their potential problems in mispecifying significance levels.

1.4.3 Tests for random walks

Another way to test the EMH is to look at the random walk model (1.14) for log prices $X_t \equiv \log P_t$. In general we may impose an autoregressive model for the log prices:

$$X_t = \mu + \alpha X_{t-1} + \varepsilon_t. \tag{1.16}$$

To test the validity of model (1.14) is equivalent to testing the hypothesis $H_0 : \alpha = 1$ in the above model. This is a special case of the unit-root test which we will revisit again later. We introduce here the Dickey-Fuller test which in fact deals with three different cases: (i) the model (1.16) with a drift μ, (ii) the model without drift

$$X_t = \alpha X_{t-1} + \varepsilon_t, \tag{1.17}$$

and (iii) the model with both drift and a linear trend

$$X_t = \mu + \beta t + \alpha X_{t-1} + \varepsilon_t. \tag{1.18}$$

Based on observations $X_1, \cdots, X_T$, let $\widehat{\alpha}$ be the least squares estimator for α, and $\mathrm{SE}(\widehat{\alpha})$ be the standard error of $\widehat{\alpha}$. These can easily be obtained from any least-squares package. Then the Dickey-Fuller statistic is defined as

$$W = (\widehat{\alpha} - 1)/\mathrm{SE}(\widehat{\alpha}). \tag{1.19}$$

We reject $H_0 : \alpha = 1$ if W is smaller than a critical value determined by the significance level of the test and the distribution of W under H_0. The intuition behind this one-sided test may be understood as follows. This random walk test is only relevant when the evidence for $\alpha < 1$ is overwhelming. Then we reject $H_0 : \alpha = 1$ only if the statistical evidence is in favor of $H_1 : \alpha < 1$. The hypothesis H_1 implies that X_t is a stationary and causal process (see section 2.2.2 below) for models (1.16) and (1.17), and, furthermore, the changes $\{X_t - X_{t-1}\}$ is an auto-correlated process. In the context of model (1.14), this implies that the returns $r_t = \log P_t - \log P_{t-1}$ are auto-correlated and, therefore, are not white noise. When $\alpha > 1$, the process X_t is explosive, which implies $r_t = \mu + \gamma \log P_{t-1} + \varepsilon_t$ for some positive constant γ. The latter equation has little bearing in modelling real financial prices except that it can be used as a tool for modeling financial bubbles; see Phillips and Yu (2011).

The least squares estimators $\widehat{\alpha}$ may easily be evaluated explicitly. For example, under (1.16), the least-squares estimate is

$$\widehat{\alpha} = \sum_{t=2}^{T} (X_t - \bar{X}_T)(X_{t-1} - \bar{X}_{T-1}) \Big/ \sum_{t=2}^{T} (X_{t-1} - \bar{X}_{T-1})^2,$$

where

$$\bar{X}_T = \frac{1}{T-1} \sum_{t=2}^{T} X_t, \qquad \bar{X}_{T-1} = \frac{1}{T-1} \sum_{t=2}^{T} X_{t-1}.$$

Furthermore let $\widehat{\mu}$ be the least squares estimator for μ in (1.16). Then

$$\mathrm{SE}(\widehat{\alpha})^2 = \frac{1}{T-3} \sum_{t=2}^{T} (X_t - \widehat{\mu} - \widehat{\alpha} X_{t-1})^2 \Big/ \sum_{t=2}^{T} (X_{t-1} - \bar{X}_{T-1})^2.$$

Under model (1.17),

$$\widehat{\alpha} = \frac{\sum_{t=2}^{T} X_t X_{t-1}}{\sum_{t=2}^{T} X_{t-1}^2}, \qquad \mathrm{SE}(\widehat{\alpha}) = \frac{\sum_{t=2}^{T} (X_t - \widehat{\alpha} X_{t-1})^2}{(T-2) \sum_{t=2}^{T} X_{t-1}^2}.$$

There also exists the Dickey-Fuller coefficient test, which is based on the test statistic $T(\hat{\alpha} - 1)$. The asymptotic null distributions are complicated, but can be tabulated.

At significant level $\alpha = 0.05$, the critical values are -8.347 and -13.96 respectively for testing model (1.17) (without drift) and model (1.16) (with drift).

Although the Dickey-Fuller statistic is of the form of a t-statistic (see (1.19)), t-distributions cannot be used for this test as all the three models under H_0 are nonstationary (see section 2.1 below). In fact the Dickey-Fuller test statistic admits certain non-standard asymptotic null distributions and those distributions under models (1.16) – (1.18) are different from each other. Fortunately the quantiles or critical values of those distributions have been tabulated in many places; see, e.g. Fuller (1996). Table 1.2 lists the most frequently used critical values, evaluated by simulation with the sample size $T = 100$. Larger sample sizes will result in critical values that are slightly smaller in absolute value and smaller sample sizes will result in somewhat larger critical values.

Table 1.2 The critical values of the (augmented) Dickey-Fuller test

Model	Significance level		
	10%	5%	1%
(1.17) or (2.66): no drift, no trend	−1.61	−1.95	−2.60
(1.16) or (2.67): drift, no trend	−2.25	−2.89	−3.51
(1.18) or (2.68): drift & trend	−3.15	−3.45	−4.04

The *R*-code "aDF.test.r" defines a function `aDF.test` which implements the (augmented) Dickey-Fuller test: `aDF.test(x, kind=i, k=0)`, where `x` is a data vector, and `i` should be set at 2 for model (1.16), 1 for model (1.17), and 3 for model (1.18).

We now apply the Dickey-Fuller test to the log daily, weekly, and monthly prices displayed in Figure 1.3. Since the returns (i.e. the differenced log prices) fluctuate around 0 and show no linear trend, we tend to carry out the test based on either model (1.16) or model (1.17). But for the illustration purpose, we also report the tests based on model (1.18). The P-values of the test with the three models for the daily, weekly, and monthly prices are listed below.

Model used	(1.17)	(1.16)	(1.18)
daily	> 0.9	0.392	0.646
weekly	> 0.9	0.336	0.698
monthly	> 0.9	0.413	0.791

Since none of those tests are statistically significant, we cannot reject the hypothesis that the log prices for the S&P 500 are random walk. This applies to daily, weekly, and also monthly data. We also repeat the above exercise for the daily, weekly and monthly returns (i.e. the differenced log prices), obtaining the P-values smaller than 0.01 for all the cases. This shows that the returns are not random walks across difference frequencies.

The Dickey-Fuller test was originally proposed in Dickey and Fuller (1979). It has been further adapted in handling the situations when there are some autoregressive terms in models(1.16) – (1.18); see section 2.8.2 below.

1.4.4 Ljung-Box test and Dickey-Fuller test

Both Ljung-Box and Dickey-Fuller tests can be used to validate some aspects of efficient markets hypothesis. First of all, the input of Ljung-Box is the returns of assets, whereas the Dickey-Fuller test utilizes the log-prices. Secondly, the alternative hypothesis of Ljung-Box is nonparametric, which merely requires the stationary correlated processes, whereas the Dickey-Fuller test is designed to test against the parametric alternative hypothesis, which is a stationary AR(1) process. Thus, the Ljung-Box test is more omnibus whereas the Dickey-Fuller is more specific. Putting both in terms of asset returns, the null hypotheses of both problems are the same: returns behaves like uncorrelated white noise. However, the alternatives of the Ljung-Box is larger, as shown in Figure 1.12.

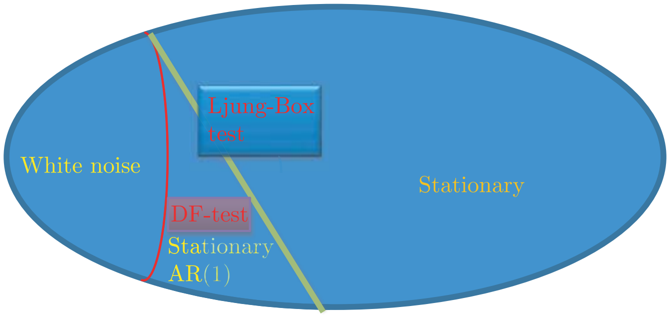


Figure 1.12 In terms of returns, the null hypotheses of both Ljung-Box and Dickey-Fuller tests are the same. However, the alternative of Ljung-Box is larger.

1.5 Appendix: Q-Q plot and Jarque-Bera test

1.5.1 Q-Q plot

A Q-Q plot is a graphical method for comparing two probability distribution functions based on their quantiles. It is particularly effective to reveal the differences in the tail-heaviness of the two distributions.

For any probability distribution function F and $\alpha \in (0, 1)$, the α-th quantile of F is defined as

$$F^{-1}(\alpha) = \max\{x : F(x) \leqslant \alpha\}. \tag{1.20}$$

For any two probability distribution functions F and G, the quantile-quantile plot, or simply the *Q-Q plot*, of F and Q is a curve on a two-dimensional plane obtained by plotting $F^{-1}(\alpha)$ against $G^{-1}(\alpha)$ for $0 < \alpha < 1$.

It can be shown that if one distribution is a location-scale transformation of the other, i.e.

$$F(x) = \sigma^{-1} G\left(\frac{x-\mu}{\sigma}\right)$$

for some constant μ and $\sigma > 0$, their Q-Q plot is a straightline. (In fact the inverse is also true.) Hence it is useful to draw a straight line passing through the two inter-quarters of the quantiles to highlight the differences in the tails of the two distributions.

We illustrate the usefulness of a Q-Q plot by an example using the daily S&P 500 returns. The lower-left panel in Figure 1.5 is the Q-Q plot of F and G, where F is the standard normal distribution and G is the empirical distribution of the daily returns of S&P 500 index. It gives basically the scatter plot of pairs

$$\left(F^{-1}\left(\frac{i-0.5}{n}\right), x_{(i)}\right), \qquad i = 1, \cdots, n,$$

where $x_{(i)}$ is the i^{th} smallest value of the data $\{x_i\}_{i=1}^n$, representing the empirical i/n-quantile, and $F^{-1}((i-0.5)/n)$ is its corresponding theoretical quantile modulus a location-scale transform. We do not use $F^{-1}(i/n)$ as its theoretical quantile to avoid $F^{-1}(i/n) = \infty$ for $i = n$. Different software has slightly different modifications from what is presented above, but the key idea of comparing the empirical quantiles with those of their referenced distribution remains the same. The blue straight line marks the position if the two distribution are identical under a location-scale transformation. The points on the left in the graph are the lower quantiles, corresponding to α close to 0 (see (1.20) above). Since those points are below the blue line, the lower quantiles of G (empirical quantiles) are smaller (i.e. negatively larger) than their expected values if G is a location-scale transformation of F. This means that the left tail of G is heavier than that of F, namely the daily returns of S&P 500 index have a heavier left tail than the normal distribution. Similarly, it can be concluded that the daily returns of S&P 500 index have a heavier right tail than the normal distribution. However, the daily returns of S&P 500 index have lighter tails than the t-distribution with degree of freedom 3, as shown in Figure 1.9.

Q-Q plots may be produced using the R-functions `qqplot, qqline` or `qqnorm`.

1.5.2 Jarque-Bera test

Q-Q plots check normality assumptions by informal graphical inspection. They are particularly powerful in revealing tail behavior of data. Formal tests can also be constructed. For example, one can employ the Kolmogorov-Smirnov test for testing normality. A popular test for normality is the Jarque-Bera test, which is defined as

$$\mathrm{JB} = \frac{n}{6}[S^2 + (K-3)^2/4]$$

where for a given sequence of data $\{x_i\}_{i=1}^n$,

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3/n}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2/n\right)^{3/2}}$$

is the *sample skewness* and

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4/n}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2/n\right)^{2}}$$

is the *sample kurtosis*. Therefore, the JB-statistic validates really only the skewness and kurtosis of normal distributions.

Under the null hypothesis that the data are drawn independently from a normal distribution, the asymptotic null distribution of the JB-statistic follows approximately χ_2^2-distribution. Therefore, the P-value can easily be computed by using χ_2^2-distribution.

1.6 Further reading and software implementation

The books that have strong impact on our writing are Fan and Yao (2003) and Campbell et al. (1997). The former emphasizes advanced theory and methods on nonlinear time series and has influenced our writing on the time series aspect. The latter emphasizes on the economic interpretation of econometric results of financial markets and has shaped our writing on the cross-sectional aspect of the book. Since preparing the first draft of the lecture notes on Financial Econometrics taught at Princeton University in 2004, a number of books on the subject have been published. For an introduction to financial statistics, see Ruppert (2004, 2010), Carmona (2004, 2013), and Franke, et. al (2015). Tsay (2010, 2013) provide an excellent and comprehensive account on the analysis of financial time series. Gourieroux and Jasiak (2001) provide an excellent introduction to financial econometrics for those who are already familiar with econometric theory. For the financial econometrics with emphasis on investments, see Rachev et al. (2013).

Most of the computation in this book was carried out using the software package R, which is free and publicly available. See Section 2.10 for an introduction and installation. The books by Ruppert (2010), Carmona (2013) and Tsay (2013) are also implemented in R.

It is our hope that readers will be stimulated to use the methods described in this book for their own applications and research. Our aim is to provide information in sufficient detail so that readers can produce their own implementations. This will be a valuable exercise for students and readers who are new to the area. To assist this endeavor, we have placed all of the data sets and codes used in this book on the following web site:

`http://orfe.princeton.edu/~jqfan/fan/FinEcon.html`

1.7 Exercises

1.1 Download the daily, weekly and monthly prices for the Nasdaq index and the IBM stock from *Yahoo!Finance*. Reproduce Figures 1.3 – 1.8 using the Nasdaq index and the IBM stock data instead.

1.2 Consider a path dependent payoff function $Y_t = a_1 r_{t+1} + \cdots + a_k r_{t+k}$ where $\{a_i\}_{i=1}^k$ are given weights. If the return time series is weak stationary in the sense that $\text{cov}(r_t, r_{t+j}) = \gamma(j)$. Show that

$$\text{var}(Y_t) = \sum_{i=1}^k \sum_{j=1}^k a_i a_j \gamma_{i-j}.$$

A natural estimate of this variance is the following substitution estimator:

$$\widehat{\text{var}}(Y_t) = \sum_{i=1}^k \sum_{j=1}^k a_i a_j \hat{\gamma}_{i-j},$$

where $\hat{\gamma}_{i-j}$ is defined by (1.8). Show that $\widehat{\text{var}}(Y_t) \geqslant 0$.

1.3 Consider the following quote from Eugene Fama who was Myron Scholes' thesis adviser:

If the population of prices changes is strictly normal, on the average for any stock $\cdots$ an observation more than five standard deviations from the mean should be observed about once every 7000 years. In fact such observations seem to occur about once every three or four years.

(See Lowenstein, 2001, page 71.) For $X \sim N(\mu, \sigma^2)$, $P(|X - \mu| > 5\sigma) = 5.733 \times 10^{-7}$, deduce how many observations per year Fama was implicitly assuming to be made. If a year is defined as 252 trading days and daily returns are normal, how many years is it expected to take to get a 5 standard deviation event? How does the answer to the last question change when the daily returns follow the t-distribution with 4 degrees of freedom.

1.4 Is the (marginal) distribution of log-returns over a long time horizon (e.g. monthly or quarterly) close to normal? Explain briefly.

1.5 Generate a random sample of size 1000 from the t-distribution with ν degrees of freedom and another random sample of size 1000 from the standard normal distribution. Apply the Kolmogorov-Smirnov test to check if they come from the same distribution. Report the results for $\nu = 5$, 10, 15 and 20.

1.6 Report the P-values for appying the Jarque-Bera test to the data given in Exercise 1.1. What can you conclude based on these P-values?

1.7 Generate a random sample of size 100 from the t-distribution with ν degrees of freedom for $\nu = 5$, 15 and ∞ (i.e. normal distribution). Apply the Jarque-Bera test to check the normality and report the P-values.

1.8 According to the efficient market hypothesis, is the return of a portfolio predictable? Is the volatility of a portfolio predictable? State the most appropriate mathematical form of the efficient market hypothesis.

1.9 If the Ljung-Box test is employed to test the efficient market hypothesis, what null hypothesis is to be tested? If the autocorrelation for the first 4 lags of the monthly log-returns of the S&P 500 is

$$\hat{\rho}(1) = 0.2, \hat{\rho}(2) = -0.15, \hat{\rho}(3) = 0.25, \hat{\rho}(4) = 0.12$$

based on past 5 years data, is the efficient market hypothesis reasonable?

1.10 Generate 400 time series from the independent Gaussian white noise $\{r_t\}_{t=1}^T$ with $T = 100$. Compute

$$Z = \sqrt{T}\hat{\rho}(1), \quad Q_m, \quad Q_m^*$$

for $m = 3, 6$, and 12. Plot the histograms of Z, Q_3, Q_3^* and Q_6 and compare it with their asymptotic distributions. Report the first, fifth and tenth percentiles of the statistic $|Z_1|$, Q_3, Q_3^*, Q_6, Q_6^*, Q_{12} and Q_{12}^*, among 400 simulations and compare them with their theoretical (asymptotic) percentiles.

1.11 Repeat the experiment in Exercise 1.10 when $T = 400$ and r_t is generated from the t-distribution with degree of freedom 5.

1.12 What is the alternative hypothesis of the Dickey-Fuller test for the random walk? Suppose that based on last 120 quarterly data on the US GDP, it was computed that $\hat{\alpha} = 0.95$. Does the US GDP follow a random walk with a drift? Answer the question at 5% significance level (the critical value is -13.96) using the Dickey-Fuller coefficient test for the model with drift.

1.13 (*Implication of martingale hypothesis*). Let S_t be the price of an asset at time t. One version of the EMH assumes that the prices of any asset form a martingale process in the sense that

$$E(S_{t+1}|S_t, S_{t-1}, \cdots) = S_t, \quad \text{for all } t.$$

To understand the implication of this assumption, we consider the following simple investment strategy. With initial capital C_0 dollars, at the time t we hold α_t dollars in cash and β_t shares of an asset at the price S_t. Hence the value of our investment at time t is $C_t = \alpha_t + \beta_t S_t$. Suppose that our investment is self-financing in the sense that

$$C_{t+1} = \alpha_t + \beta_t S_{t+1} = \alpha_{t+1} + \beta_{t+1} S_{t+1},$$

and our investment strategy $(\alpha_{t+1}, \beta_{t+1})$ is entirely determined by the asset prices up to the time t. Show that if $\{S_t\}$ is a martingale process, there exist no strategies such that $C_{t+1} > C_t$ with probability 1.

Chapter 2

Linear Time Series Models

Data obtained from observations collected sequentially over time are common in this information age. For example, we have collections on daily stock prices, weekly interest rates, monthly sales figures, quarterly consumer price indices (CPI) and annual gross domestic product (GDP) figures. Those data collected over time are called time series. The purpose of analyzing time series data is in general two-fold: to understand the stochastic mechanism that generates the data, and to predict or forecast the future values of a time series. This chapter introduces a class of linear time series models, or more precisely, a class of models which depict the linear features (including linear dependence) of time series. Those linear models and associated inference techniques provide the basic framework for the study of the linear dynamic structure of financial time series and for forecasting future values based on linear dependence structures.

2.1 Stationarity

One of the important aspects of time series analysis is to use the data collected in the past to forecast the future. How can historical data be useful for forecasting a future event? This is through the assumption of stationarity which refers to some time-invariance properties of the underlying process. For example, we may assume that the correlation between the returns of tomorrow and today is the same as those between any two successive days in the past. This enables us to aggregate the information from the data in the past to learn about the correlation. This correlation invariance over time is a typical characteristic of the so called *weak stationarity* or *covariance stationarity* . It facilitates linear prediction which is essentially based on the correlation between a predicated variable and its predictor (such as in linear regression). A stronger time-invariant assumption is that the joint distribution of the returns in a week in the future is the same as that in any weeks in the past. In other words, prediction is always based on some invariance properties over time, although the invariance may refer to some characteristics of the probability distribution of the process, or to the law governing the change of the distribution. We introduce the concept of stationarity more formally below.

A sequence of random variables $\{X_t,\ t = 0, \pm 1, \pm 2, \cdots\}$ is called a stochastic process and is served as a model for a set of observed time series data. It is convenient to refer to $\{X_t\}$ itself as a time series. It is known that the complete probability structure of $\{X_t\}$ is determined by the set of all the finite-dimensional distributions of $\{X_t\}$. Fortunately most linear features concerned depend on the first two moments of $\{X_t\}$, which are the main objects depicted in linear time series models. Of course if $\{X_t\}$ is a Gaussian process in the sense that all its finite-dimensional distributions are normal, the first two moments then determine the the probability structure of $\{X_t\}$ completely and $\{X_t\}$ is a linear process.

A time series $\{X_t\}$ is said to be *weakly stationary* (or *second order stationary* or *covariance stationary*) if $E(X_t^2) < \infty$ and both EX_t and $\text{cov}(X_t, X_{t+k})$, for any integer k, do not depend on t.

For weakly stationary time series $\{X_t\}$, let $\mu = EX_t$ denote its common mean. We define the *autocovariance function* (ACVF) as

$$\gamma(k) = \text{cov}(X_t, X_{t+k}) = E\{(X_t - \mu)(X_{t+k} - \mu)\}, \tag{2.1}$$

and the *autocorrelation function* (ACF) as

$$\rho(k) = \text{Corr}(X_t, X_{t+k}) = \gamma(k)/\gamma(0) \tag{2.2}$$

for $k = 0, \pm 1, \pm 2, \cdots$. Note that $\gamma(0) = \text{var}(X_t)$ is independent of t. For simplicity, we drop the adverb "weakly" and call $\{X_t\}$ *stationary* if it is weakly stationary, i.e. $\{X_t\}$ has finite and time-invariant first two moments. It is easy to see that $\rho(0) = 1$ and $\rho(k) = \rho(-k)$ for any stationary processes, and that the variance-covariance matrix of the vector $(X_t, \cdots, X_{t+k})$ is

$$\text{var}(X_t, \cdots, X_{t+k}) = \begin{pmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \cdots & \gamma(k-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & \cdots & \gamma(k-2) \\ \vdots & \vdots & \vdots & & \vdots \\ \gamma(k-2) & \gamma(k-3) & \gamma(k-4) & \cdots & \gamma(1) \\ \gamma(k-1) & \gamma(k-2) & \gamma(k-3) & \cdots & \gamma(0) \end{pmatrix}.$$

Therefore, for any linear combinations,

$$\begin{aligned} \text{var}(\sum_{i=1}^k a_i X_{t+i}) &= \sum_{i=1}^k \sum_{j=1}^k a_i a_j \text{cov}(X_{t+i}, X_{t+k}) \\ &= \sum_{i=1}^k \sum_{j=1}^k a_i a_j \gamma(i-j) \geqslant 0. \end{aligned} \tag{2.3}$$

As such, the function $\gamma(\cdot)$ is referred to as the *semi-positive definite* function.

A very specific class of processes that plays a similar role to zero in the number theory is the *white noise*. When $\rho(k) = 0$ for any $k \neq 0$, $\{X_t\}$ is called a *white noise*, and is denoted by $X_t \sim \mathrm{WN}(\mu, \sigma^2)$, where $\sigma^2 = \gamma(0) = \mathrm{var}(X_t)$. In other words, a white noise is a sequence of uncorrelated random variables with the same mean and the same variance. White noise processes are building blocks for constructing general stationary processes.

In practice we use an observed sample $X_1, \cdots, X_T$ to estimate ACVF and ACF by the *sample ACVF* and *sample ACF*. They are basically the sample covariance and sample correlation of the lagged pairs $\{(X_{t-k}, X_k)\}_{t=k+1}^T$. Formally, they are defined as follows

$$\widehat{\gamma}(k) = \frac{1}{T} \sum_{t=k+1}^{T} (X_t - \bar{X})(X_{t-k} - \bar{X}), \qquad \widehat{\rho}(k) = \widehat{\gamma}(k)/\widehat{\gamma}(0), \tag{2.4}$$

where $\bar{X} = T^{-1} \sum_{1 \leqslant t \leqslant T} X_t$. In the estimator $\widehat{\gamma}(k)$, we use the divisor T instead of $T - k$. This is a common practice adopted by almost all statistical packages. It ensures that the function $\hat{\gamma}(\cdot)$ is semi-positive definite (Exercise 2.2), a property given by (2.3). See Fan and Yao (2003) pp.42 for further discussion on this choice.

Weak stationarity is indeed a very weak notion of stationarity. For example, if $\{X_t\}$ is weakly stationary, it does not imply that $\{X_t^2\}$ is weakly stationary. Yet, the latter time series has very strong connections with the volatility of financial returns. Therefore, we need a stronger version of stationarity as follows.

A time series $\{X_t, \, t = 0, \pm 1, \pm 2, \cdots\}$ is said to be *strongly stationary* or *strictly stationary* if the k-dimensional distribution of $(X_1, \cdots, X_k)$ is the same as that of $(X_{t+1}, \cdots, X_{t+k})$ for any $k \geqslant 1$ and t.

This assumption is needed in the context of nonlinear prediction. Obviously the strict stationarity implies the (weak) stationarity provided $E(X_t^2) < \infty$. In addition, the strong stationarity of $\{X_t, \, t = 0, \pm 1, \pm 2, \cdots\}$ implies that of the time series $\{g(X_t), \, t = 0, \pm 1, \pm 2, \cdots\}$ is also strongly stationary for any function g.

2.2 Stationary ARMA models

One of the most frequently used time series models is the *stationary autoregressive moving average* (ARMA) model. It is frequently used in modeling the dynamics on returns of financial assets and other time series.

2.2.1 Moving average processes

Perhaps the simplest stationary time series are moving average (MA) processes. They also facilitate the computation of the autocovariance function easily. A simple example of this is the k-period return (1.6).

Let $\varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$. For a fixed integer $q \geqslant 1$, we write $X_t \sim \mathrm{MA}(q)$ if X_t is defined as a moving average of q successive ε_t as follows:

$$X_t = \mu + \varepsilon_t + a_1\varepsilon_{t-1} + \cdots + a_q\varepsilon_{t-q}, \tag{2.5}$$

where μ, $a_1, \cdots, a_q$ are constant coefficients. In the above definition, ε_t stands for the innovation at time t, and the innovations $\varepsilon_t, \varepsilon_{t-1}, \cdots$ are unobservable. The intuition behind the moving average equation (2.5) may be understood as follows: innovation ε_t stands for the shock to the market at time t and X_t is the impact on the return from the innovations up to time t. The coefficient a_k is regarded as a "discount" factor on the k-lagged innovation ε_{t-k}. For example, for $a_k = b^k$ and $|b| < 1$, the impact of ε_t fades away exponentially over time.

In fact X_t defined by (2.5) is always stationary with $EX_t = \mu$, as the coefficients $a_1, \cdots, a_q$ do not vary over time. We first use a simple example to illustrate how to calculate ACVF and ACF for MA processes.

Example 2.1 *For MA(1) model $X_t = \mu + \varepsilon_t + a\varepsilon_{t-1}$, it holds*

$$\gamma(0) = \mathrm{var}(X_t) = \mathrm{var}(\varepsilon_t) + \mathrm{var}(a\varepsilon_{t-1}) + 2\mathrm{cov}(\varepsilon_t, a\varepsilon_{t-1}) = (1 + a^2)\sigma^2.$$

Similarly,

$$\gamma(1) = \mathrm{cov}(X_t, X_{t-1}) = \mathrm{cov}(\varepsilon_t + a\varepsilon_{t-1}, \varepsilon_{t-1} + a\varepsilon_{t-2}) = a\sigma^2,$$

since there is only one common term, ε_{t-1} in both X_t and X_{t-1}. Now for the ACVF of lag two, we have

$$X_t = \mu + \varepsilon_t + a\varepsilon_{t-1},$$

which does not share the same set of innovations $\{\varepsilon_t\}$ as

$$X_{t-2} = \mu + \varepsilon_{t-2} + a\varepsilon_{t-3}.$$

Therefore $\gamma(2) = 0$. Similarly, $\gamma(k) = 0$ for any $k \geqslant 2$. Consequently,

$$\rho(1) = a/(1 + a^2), \quad \rho(k) = 0 \text{ for any } |k| > 1. \tag{2.6}$$

Since $2|a| < 1 + a^2$, $|\rho(1)| \leqslant 0.5$ for any MA(1) process.