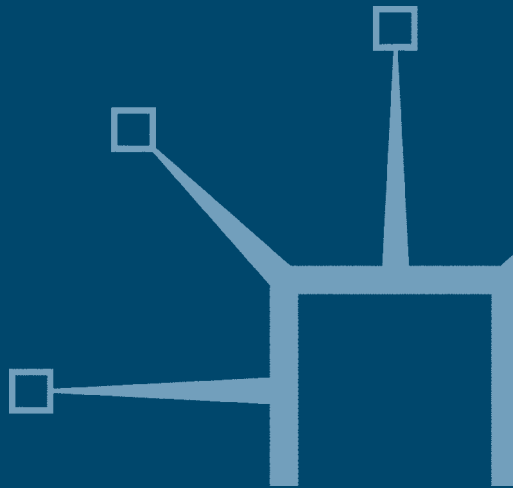# Energy Risk Modeling

Applied Modeling Methods for Risk Managers

Nigel Da Costa Lewis

ENERGY RISK MODELING

This page intentionally left blank

Energy Risk Modeling

Applied Modeling Methods for Risk Managers

NIGEL DA COSTA LEWIS

palgrave
macmillan

In honor of my Father, Merton Da Costa Henry. His inquiring mind and hilarious tales bring smiles to all around him

This page intentionally left blank

Contents

List of Tables

List of Figures

Preface

One of the most difficult problems encountered by an energy risk manager, who may not be very statistical, is the lack of an instructive text which can provide an intuitive interface between the concepts and tools necessary for acquiring competence in the applied statistical modeling side of energy risk management. This text is written primarily for those individuals and also for those who need a gentle introduction in how to go about modeling energy price risk. A secondary purpose in writing this text is to bring my approach to the art of applied statistical modeling to people who are not necessarily versed in applied statistical thinking – actuaries, managers, traditional risk analysts and risk managers, product consultants, and sales people. Whilst these individuals are capable in their fields, they are not necessarily familiar with the fundamentals or particulars of applied statistical reasoning.

In the years since financial risk management emerged as a separate discipline from the more general field of Finance and Economics, the modeling techniques available have undergone steady improvement. However, newcomers to the field of energy price risk management still lack an elementary introduction which allows them to make speedy progress in acquiring the knowledge required to take full advantage of these emerging methods. Key to success is a thorough understanding of the art and science of statistical methods. With this in mind, this text has been written to provide a rapid primer on energy price risk modeling illustrated through elementary and more advanced statistical methods. In this sense, it is an introduction to energy price risk modeling embedded in a practical guide for using and understanding applied statistical methods.

It is quite possible to write a reference text on this subject, in terms of mathematical assumptions, lemmas, proofs, and conjectures. However, such texts are daunting to both the newcomer and practicing energy risk manager who may simply require an easy-to-follow guide which enhances their understanding whilst at the same time encouraging creativity. It is for this reason

that I have decided, in this text, not to describe energy risk modeling in terms of abstract mathematical concepts but rather as an applied living art embedded in the science of statistical and mathematical methods.

Recognizing that the successful mastering of applied modeling depends heavily on the working of practical examples and problems I have included a large number of review questions at the end of each chapter. Working carefully through each one will reinforce many of the ideas discussed and enhance the reader's understanding of the notions discussed. Almost always modeling is carried out using specialized software on a computer. To aid the reader I have included a number of examples of code written in *R*, a numerical and statistical computing package. The samples can be easily ported to other mathematical packages and spreadsheet programs. I have chosen to use *R* in this book because it is free (downloadable from r-project.org), very easy to use, and can run on all the popular computing operating systems. Naturally, any such compilation as this must omit some models and methods. In choosing the material I have been guided by my own personal interest, comments by users of my research boutique StatMetrics, and by the views of my colleagues and friends. As always with my books lucidity of style and simplicity of expression have been my twin objectives.

Nigel Da Costa Lewis

This page intentionally left blank

The Statistical Nature of Energy Risk Modeling

In O. Henry's *The Handbook of Hymen*, Mr Pratt is wooing the wealthy Mrs Sampson. Unfortunately for Mr Pratt, he has a rival – a romantic poet. In order to compensate for his romantic disadvantage, the studious Mr Pratt selects quotes from a text on statistical facts in an audacious attempt to dazzle Mrs Sampson into marrying him:

> "Let us sit on this log at the roadside," says I, "and forget the inhumanity and ribaldry of the poets. It is in the glorious columns of ascertained facts and legalized measures that beauty is to be found. In this very log we sit upon, Mrs. Sampson," says I, "is statistics more wonderful than any poem. The rings show it was sixty years old. At the depth of two thousand feet it would become coal in three thousand years. The deepest coal mine in the world is at Killingworth, near Newcastle. A box four feet long, three feet wide, and two feet eight inched deep will hold one ton of coal. If an artery is cut, compress it above the wound. A man's leg contains thirty bones. The Tower of London was burned in 1841. "Go on, Mr Pratt," says Mrs Sampson. "Them ideas is so original and soothing. I think statistics are just lovely as they can be."

Many traditional risk managers, unlike Mr Pratt and Mrs Sampson have had a disdain for quantitative methods, and statistical methods in particular. In fact the importance of statistical science in risk management is only now beginning to be recognized by the energy risk management profession. Stories of statistically illiterate risk mangers are legion. Immediately after completing my PhD at Cambridge University I was interviewed by a head of global risk for a large financial institution who wondered aloud on the value of employing a statistician to monitor risk. Out of date and out of touch, the

individual clearly had not absorbed the widely accepted wisdom:

> Not only can statistical analysis be used to identify relationships, but processes can be discovered that actually function according to statistical principles...more and more, researchers are discovering that the principles of statistics are not only useful for identifying relationships between sets of empirical data – i.e., worldly observations – but also for the theoretical modeling of the process or processes that generate the data. In other words, instead of the application of statistical principles as a means to an end, the principles become an end in themselves, as they serve to explain aspects of our existence.[1]

Alas, that particular individual who interviewed me is no longer the head of global risk for the organization in question, or any other organization. Why? Because a risk manager who does not understand how risk is calculated or modeled is like an accountant who lacks knowledge of how the figures which make up a financial statement were derived. Such individuals may well be able to offer a limited interpretation but are unable to assess or defend the assumptions made in arriving at a figure. Fortunately, in accountancy, professional bodies with legal mandates enforce minimum standards thus preventing unqualified individuals from practicing. In energy risk management there are no such requirements. Whilst statistically illiterate risk managers still exist they are finding it increasingly difficult to maintain their status as risk professionals without some training in statistical and quantitative methods.

This book provides a solid foundation for those seeking some insight into the statistical aspects of modeling energy risk. It departs from other texts in this field by its unique emphasis on the process that underlies the art and science inherent in statistical modeling rather than the pedantic details of theoretical finance or mathematics. Through careful reading it is hoped that the reader will gain some understanding of how to utilize statistical thinking in defining and solving problems associated with energy risk management. Wherever possible words are used as explanation rather than equations and the focus is always on quantitative reasoning and understanding rather than mathematical conjectures, lemmas, and proofs. Before beginning our journey, it is instructive to say a little about the history and evolution of energy.

1.1   HISTORICAL EVOLUTION OF ENERGY

The use of energy products by mankind dates back to approximately 500,000 BC when the discovery of fire allowed early humans to exert some control over the environment. Wood and dried animal dung were primary sources of fuel; their use ranged from heating and cooking to lighting, fighting and frightening away predatory animals. Human cultivation of wood as a fuel source and for use in manufacturing dates back over 50,000 years.

By around 3500 BC simple machines, such as ramps, levers, pulleys, and the wheel allowed man to begin substituting machinery for human and animal muscle power. Indeed, the stone circle at Stonehenge in England, the Pyramids of ancient Egypt, and the Great Wall of China are ancient testimony to the ongoing process of the substitution of machinery for human and animal labor.

The Egyptians made the earliest known sailboats around 3500 BC. However, it was not until 700 BC that wind energy found more widespread use, for example to separate the chaff from grains such as wheat and rice, and to power huge boats on journeys of discovery across vast oceans and seas. Within a couple of hundred years, by 500 BC, the water-based technologies of the ancient Greeks and Romans (used for driving machinery such as millstones) had been combined with wind to give mankind access, through the use of windmills and paddlewheels, to a more reliable source of energy.

By 100 AD the burning of coal for heat, cooking, and the manufacture of metals provided an increasingly important substitute for wood. Its higher heat content per unit of weight was an important advantage. During this period, wood, water, and wind were the major sources of energy. By 640 AD the Persians had perfected the design of windmills used to grind grain. Whilst in the 1300s early German engineers designed wood-burning blast furnaces allowing them to produce vast quantities of iron. However, it was not until the late 1600s with the deforestation of England that we see the use of coal beginning to dominate other energy sources with the invention of steam engines. Indeed, the engineer Thomas Savery invented one of the earliest steam engines in response to constant flooding of English coal mines. In 1698 he obtained a patent for his invention named "Miner's friend" for "suck up the water with the fire." Despite the fact that the Miner's friend was little used and soon abandoned, steam engines were soon powering factories, farm implements and mass transportation. For example, in 1783, improved steam engine design resulted in the French producing the first working paddle wheel steamboat. Not long after, the people of Cornwall in southern England began using gas made from coal to fuel street lamps. More convenient than wind, water, or wood and less expensive than a stable full of horses or oxen, by 1799, coal-fired steam engines provided the backbone for the industrial revolution and the rise of British and European industry. It was thus during this century that coal and steam replaced wind and animals in sea and land transportation and wood and water in manufacturing.

Whilst the technological advances in energy production and use between 1700–99 were quite astounding, the 1800s saw the rise of a new form of energy – electricity. As 1800 dawned Alessandro Volta published a description of a silver/zinc battery. A few years later in 1807 Humphrey Davy constructed a practical battery. By 1831 Michael Faraday had built the first electric generator which converted rotary mechanical power into electric power. A little over 50 years from the first Faraday electric generator

Thomas Edison in 1882 had installed three generators at Pearl Street Station in New York. It fed power to around 5000 lamps in 225 houses. By l899, mass transportation using steam trains fired by coal and coal-fired electricity generating stations were key sources of energy for the leading European and North American cities. During this period natural gas consumption also rose dramatically. Although around 200 BC the Chinese had used natural gas-fired evaporators to produce salt from salt water. It was demand from growing European and North American populations for lighting which lead to a rapid expansion in its consumption.

With the development of motorized transportation in the early 1900s came the need for liquid sources of energy. Petroleum-based products, first marketed widely in the 1860s for lubrication of machinery and as a substitute to whale oil, proved ideal for motorized vehicles. Mass production of motor cars created a huge demand for gasoline. By 1920 there were more than 9 million motor vehicles in the United States alone. Within 30 years by 1950, petroleum-based products had become the most widely used energy source in the history of man kind.

Beginning in the 1970s energy market deregulation spurred a global trend toward the complete commoditization of energy products. Both Europe and North America witnessed a rapid growth in financial contracts for various energy products and a rapid rise in the number of nontraditional energy companies moving into the energy market place. The rapid commoditization of energy markets, especially crude oil, electricity, and natural gas, has elevated the importance of real time information, reduced the product cycle, made services fungible, narrowed margins and contributed to increased price volatility. Alongside this, the consumption of petroleum continued to rise throughout the remaining years of the twentieth century and into the twenty-first century. Fears of instability in the supply of oil have precipitated considerable interest in nuclear, hydroelectric, and solar energy sources. However, crude oil still dominates the energy sector.

1.2   FINANCIAL RISKS OF ENERGY

Risk is everywhere in the business environment. Managers and business leaders accept the relationship between risk and reward. The higher the risk to an organization the larger the potential rewards. But what exactly is risk? For our purposes we shall assume it is synonymous with uncertainty. For example, a producer of crude oil may be concerned about the uncertainty of the price of crude oil three months from now. The concern may be that a general decline in price will have serious implications for the revenue stream. A refiner on the other hand may be concerned about future price rises implying higher oil costs, which if they cannot be passed on, will result in lower revenues. Uncertainty about future energy prices we refer to as *energy*

price risk.[2] Energy price risk leaves producers and distributors exposed to unstable revenue streams and consumers exposed to unpredictable prices.

Of course players in energy markets are concerned with more than just price risk. A particular energy player may have concerns about the level of its financial exposure to another counterparty: such a concern is known as credit risk. *Credit risk* is the risk that the counterparty will default on its obligation. Energy risk players are also concerned about the liquidity of the instruments they trade. Liquidity refers to the ease and certainty by which a financial instrument can be converted into cash. A liquidity crunch occurs when bid/offer spreads for financial instruments widen out to levels that make it prohibitively expensive for players to trade. *Liquidity risk* is perhaps more of an issue in over-the-counter than exchange-traded transactions. This is because market-makers for exchange-traded instruments are required to show prices to their customers at all times. Indeed, in some markets, the size of the bid/offer spread is capped to prevent the occurrence of a liquidity crunch. However, over-the-counter transactions are more susceptible because there is no guarantee that the market-maker who sold you a particular product will make a price when the market becomes unusually volatile.[3]

Another risk important to energy market players is *basis risk*. Consider an energy participant who believes he has a fully hedged position using futures contracts. Basis risk exists because it is possible that the value of the hedge may not move exactly in line with the value of the price exposure that the participant is trying to manage. Such situations occur when there is a breakdown of the expected or usual differential between two sets (or more) of prices. Basis risk occurs partly because financial instruments do not correlate perfectly to the underlying physical markets. There is no guarantee that the price of energy futures contracts will converge to the physical markets. Theoretically they should, but in practice they may not.

Operational risk is another important risk facing energy participants. The British Bankers Association defined it as

> The risks associated with human error, inadequate procedures and control, fraudulent and criminal activities; the risk caused by technological shortcomings, system breakdowns; all risks which are not "banking" and arising from business decisions as competitive action, pricing etc; legal risk and risk to business relationships, failure to meet regulatory requirements or an adverse impact on the bank's reputation; "external factors" include: natural disasters, terrorist attacks and fraudulent activity etc. British Bankers Association (1997)

Another frequently quoted definition is that provided by the Basel Committee on Banking Supervision:

> The risk of loss resulting from inadequate or failed internal processes, people systems or from external events. Basel Committee on Banking Supervision (2001)

The key point to note is that operational risk measures the uncertainty surrounding events that cause business disruption such as failure of controls, errors, omissions, and external events such as power failures or earthquakes.

1.3 THE ROLE OF ECONOMICS – ELEMENTS OF PRICE THEORY

Traditional economic analysis of energy markets considers the effects of policy measures such as price controls and regulation. It might include an examination of a variety of social, political, legal, regulatory, environmental, and technological issues from regional, national, and even global perspectives. The objective being to inform current policy makers of the issues arising from the relationships among energy use, economic growth, and the environment. Whilst it is arguable that economic analysis can lead to a better understanding of these topics our interest in economics lies solely in the fact that economic theory provides a mechanism for price determination and thus a mechanism for analyzing price risk.

To gain insight into economic price theory we shall consider a very simple exchange economy which is closed, that is without exports and imports. Furthermore, assume the existence of $n \geq 1$ goods and $m \geq 1$ customers. Some of the n goods can be considered energy products, others may be food, telephone services and so on. For parsimony we shall also assume there is no production and each of the m customers has an initial endowment of goods given by:

$$e_k = \{q_1, q_2, ..., q_n\},$$

where $k = 1, 2, ..., m$ and q_i represents the quantity of the good i. For example, customer 2 may have an initial endowment of:

$$e_2 = \{6 \text{ units of good } 1, 70 \text{ units of good } 2, ..., 1 \text{ unit of good } n\}$$

The initial endowment confers some utility on the customer, this usefulness is captured in the individuals utility function $u(q_1, q_2, ..., q_n)$. Total utility measures the total usefulness gained from a specific endowment of goods and services. Figure 1.1 shows the total utility of a typical customer where all but one of the quantities of their initial endowment are held constant. From the diagram we see that total utility increases as the quantity of the good consumed increases. Notice however that whilst total utility is increasing it is doing so at a decreasing rate. The rate of change in total utility is known as marginal utility.
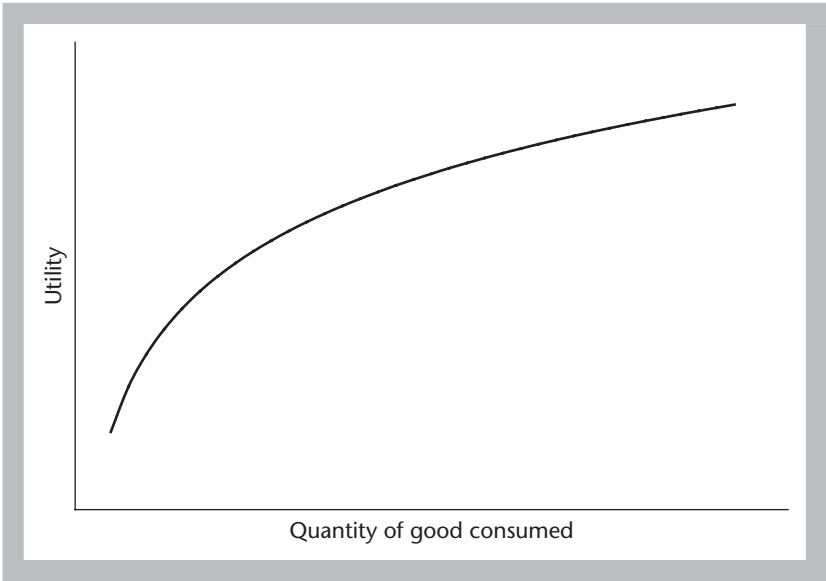
Figure 1.1 Total utility of a customer, where all but one of the quantities of their initial endowment are held constant

Marginal utility is the additional utility gained from a small additional amount of the ith good. For a particular individual it is calculated as:

$$u_i'(q_1, q_2, ..., q_n) = \frac{\partial u(q_1, q_2, ..., q_n)}{\partial q_i}$$

The law of diminishing marginal utility states that the marginal utility a good confers decreases as the quantity of the good consumed increases, ceteris paribus. In other words, whilst more of a good is deemed better than less, total utility increases at a decreasing rate as the quantity of the good consumed increases. This implies:

$$u_i'(q_1, q_2, ..., q_n) > 0$$

Given that we have assumed more is preferred to less we also know that:

$$u_i''(q_1, q_2, ..., q_n) = \frac{\partial^2 u(q_1, q_2, ..., q_n)}{\partial q_i^2} < 0$$

Therefore as shown in Figure 1.2, marginal utility is a downward sloping function.

An allocation is said to be efficient if it satisfies the pareto optimality criterion. In our example an allocation of endowments is pareto optimal

Figure 1.2 A marginal utility function of a customer, where all but one of the quantities of their initial endowment are held constant

among the *m* consumers if it is not possible to make someone better off without making someone else worse off. If we assume individuals seek to maximize their utility and if the initial distribution of endowments is not pareto optimal, the customers will start trading to increase their utility. Therefore a market price for each good will emerge. Eventually through trading, the sum of demands of the customers will equal the sum of the endowments of the customers and an equilibrium set of prices for all goods will emerge.

Classical economic theory therefore assumes that the price of a particular good and the quantity exchanged in a market are determined by the interaction of demand and supply. The demand and supply in each market is determined by the individual demand and supply of utility maximizing customers. Individual demand reflects an individual consumer's marginal utility for a particular good and translates into the quantity of a good they are willing to buy at each price. Rational individuals will adjust their consumption of a good to a level such that the marginal utility equals the cost (price) of the good. This implies the demand curve for a good is downward sloping. As shown in Figure 1.3, the demand curve depicts how much of a good a consumer is willing to buy at various prices. From the diagram we see that when the price is high a consumer is willing to buy less than when the price is low. Individual supply reflects the quantity of the good the individual agent is prepared to offer for sale at each price. It is an upward sloping curve as shown in Figure 1.4.
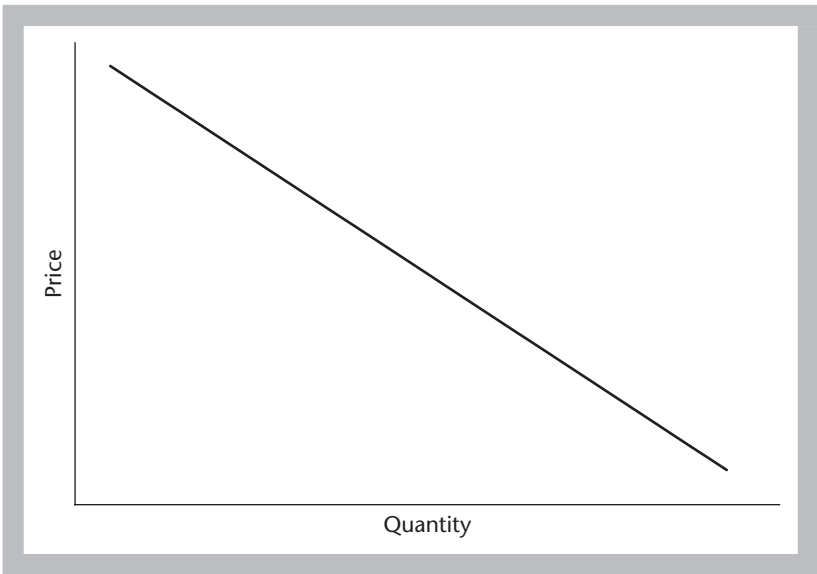
Figure 1.3 Typical demand curve
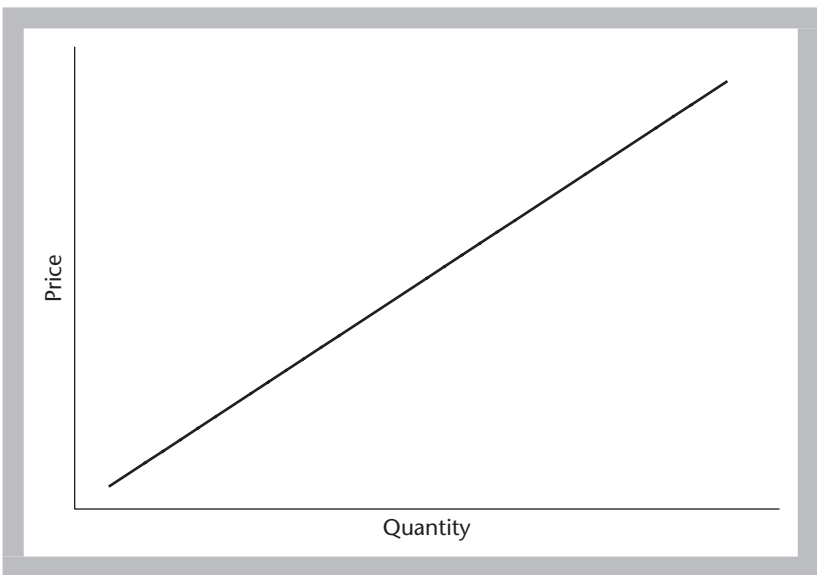
Figure 1.4 Typical supply curve

The market demand and supply for each product is therefore the sum aggregate of individual customers demands and supplies. The equilibrium price occurs where supply is equal to demand. This is depicted in Figure 1.5 where the equilibrium price is p_0 with resultant quantity of q_0. Thus we
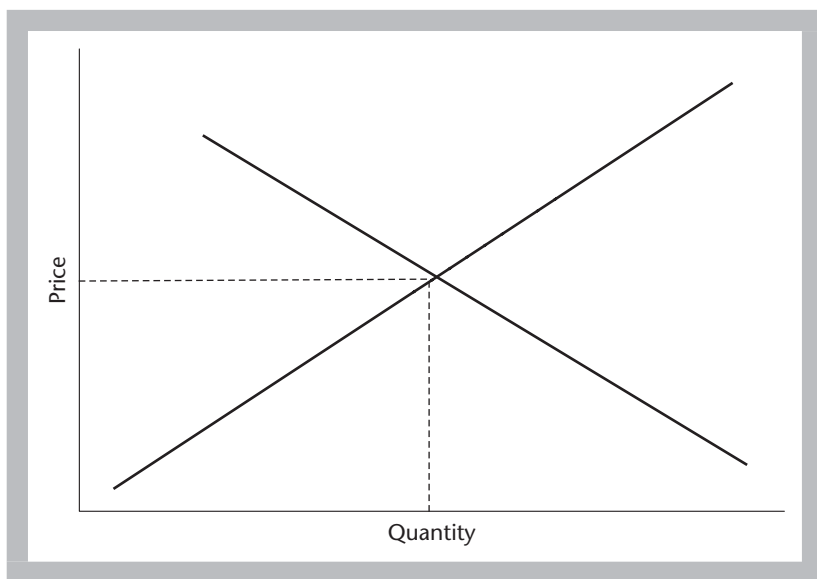
Figure 1.5 Equilibrium price and quantity

arrive at the law of supply and demand, the mechanism by which prices are determined. The law states that the price of any one good increases when the demand for that good exceeds the supply and decreases in the opposite case. The law of supply and demand is powerful because it predicts the price and quantity supplied when a market is in equilibrium. Fluctuations in price are an important part of a competitive market because they send price signals to buyers and sellers about the tightness of the market. Producers interpret these signals as either an opportunity to increase production (to meet demand) or decrease production (because there is an overabundance of supply on the market).

Example 1.1   Economic analysis of the price of crude oil

The price of crude oil is sensitive to the level of inventories. For example, in mid-December 2003 a weekly US federal report observed that crude oil inventories had declined from 277.9 million barrels to 272.8 million barrels. This was an unexpected decline, as the economy was showing signs of strong economic growth. Following the announcement the price of benchmark crude for January delivery rose to its highest level for almost a year closing at \$33.35 a barrel. This situation is illustrated diagrammatically in Figure 1.6. Strong economic growth results in an increase in demand causing the demand curve to shift from d_0 to d_1. As the demand was unexpected, production remains unchanged but inventories begin to decline, resulting
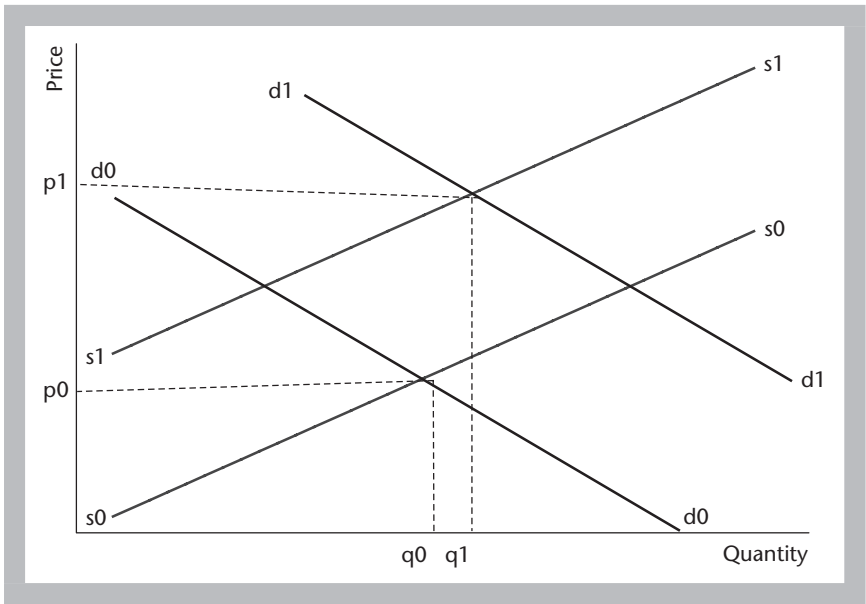
Figure 1.6 Impact of an increase in demand and reduction in supply on the crude oil spot price

in a reduction in supply and a shift of the supply curve s_0 to s_1. The immediate result is a increase in price from p_0 to p_1. The reasons for the price swings can thus been seen to be related to issues surrounding over- or under-production, production driven by factors such as recessions/booms in the world economy and severe or mild winter weather.

In practice the demand for energy is not derived directly from preferences for the energy commodity itself, rather it is derived from desires to use energy to produce a good or provide a service. For example, postal delivery workers use gasoline in order to transport goods from place a to place b. The amount of gasoline consumed by a delivery vehicle is proportional to the miles driven and inversely proportionate to the efficiency by which gasoline is converted to mechanical energy. Demand for gasoline is thus derived from choices about distances postal delivery vehicles are driven and their energy conversion efficiencies. In other words it depends primarily on demand for desired services, availability and properties of energy conversion technologies, and costs of energy and technologies used for conversion.

Nevertheless, the economic mechanism for price determination, the interaction between demand and supply remains relevant. To see that this is the case notice, for example, that in the USA the retail price of gasoline can be broken down into three components: the cost of crude oil, taxes, and a seller markup. The latter includes both a refiner margin and a retail margin. The
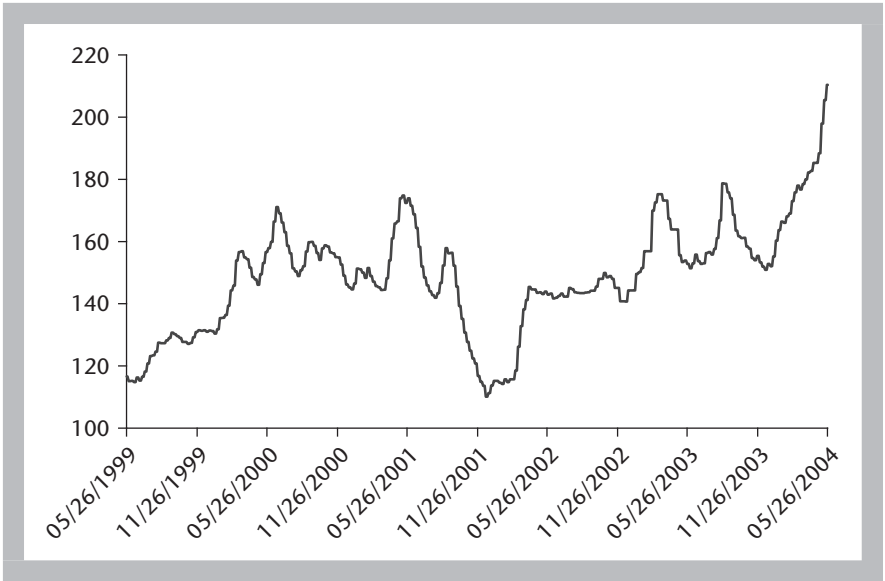
Figure 1.7 Average price of all grades of gasoline in the USA (cents per gallon)

refiner margin is the spread between the wholesale price of gasoline, the price charged by refineries to retailers, and the price of crude oil – in this case West Texas Intermediate. The retail margin is the difference between the pump price and the wholesale price of gasoline. On average, crude oil costs account for about 50% of the overall pump price, while taxes and the markup account for 30% and 20% of the overall price, respectively.

Figure 1.7 shows the average price of all grades of gasoline in the USA (cents per gallon) over the period from end of May 1999 to end of May 2004. Given the relative stability of gasoline tax rates, most of the observed short-term volatility in gasoline prices must be attributed to fluctuations in the balance between demand and supply. On the supply side the level of gasoline stocks plays a major role. A tight inventory balance exerts upward pressure on wholesale prices beyond that which can be attributed to changes in crude oil prices. This widens the spread between spot gasoline prices and crude oil prices. On the demand side, during the summer driving season demand for gasoline increases pushing prices at the pumps higher.

European markets face the same economic forces as those experienced in the USA. For example, during May 2004 prices for gasoline in Germany hit the highest levels in the country's postwar history when a liter of premium gasoline reached €1.16, the primary driver being the rising spot price of Brent Crude, which exceeded $39 per barrel, and uncertainty following political and social upheaval in the Middle East.

1.4   ENERGY MARKETS AND PRODUCTS

Energy consumption today is dominated by three products: crude, electricity, and natural gas. They are traded in a number of forms spot, futures, options, swaps, over-the-counter, and on a number of standardized exchanges such as the International Petroleum Exchange in London (IPE), New York Mercantile Exchange (NYMEX), Tokyo Commodity Exchange and the Singapore Exchange. As with all futures exchanges counterparty default risk is eliminated because the exchange acts as the counterparty to every trade.

1.4.1   Crude oil

Crude oil prices, as any other commodity, are driven by the law of supply and demand with wide price swings in times of shortage or over supply. A key determinant of supply are the actions of the members of the Organization for Petroleum Exporting Countries (OPEC), created at the Baghdad Conference during September 1960, by Iran, Iraq, Kuwait, Saudi Arabia, and Venezuela. OPEC's objective was to control the supply of crude oil onto world markets. Over the years the organization has gradually expanded to include Qatar (1961), Indonesia (1962), Libya (1962), United Arab Emirates (1967), Algeria (1969), Nigeria (1971), Ecuador (1973–92), and Gabon (1975–94).

Through a process of production quotas on member countries, OPEC collectively have the power to exert a considerable degree of influence over the price of crude oil. Perhaps the most important member of the group is Saudi Arabia partly because it accounts for around 10% of world production. Saudi Arabia is also the only producer with enough spare capacity to have a rapid effect on oil prices by increasing or decreasing production. Indeed, it is claimed that Saudi Aramco, the national oil company, can increase output by 1.5 million barrels a day within 48 hours.

Whilst in theory OPEC has the potential to control oil prices, the organization has in fact resided over a steady decline in the real value of crude oil. This is partly due to the lack of any real incentive for OPEC members to remain within their quotas and partly associated with increased efficiency of oil burning products.

Brent Crude oil is the reference for over 60% of world production: this is in a large part due to the fact that it is free from government intervention (unlike OPEC). The major markets for Brent Crude oil futures are NYMEX and the IPE. On both exchanges the contracts trade in units of 1000 barrels. The importance of oil and NYMEX for world energy markets should not be understated.

International oil is the senior energy market. It has had a colorful history of letting market forces ebb and flow. ... It is the most globalized of businesses, and at its very

epicenter is the mother of all energy markets, the New York Mercantile Exchange (NYMEX) crude oil future market. The rise to prominence of the NYMEX oil futures market is a signal achievement in the history not only of the oil business but also of the global economy. Notwithstanding the continuing prominence of the Organization of Petroleum Exporting Countries (OPEC), world oil prices today are made largely on the floor of the NYMEX exchange. Virtually all oil contracts are based on NYMEX prices, or the prices of other exchanges that are directly or indirectly tied to the NYMEX. OPEC member countries no longer try to impose fixed oil prices. Instead, they tie their prices to differentials to the NYMEX or an associated benchmark price.[4]

Energy products such as gasoline, heating oil, and jet fuel are derived from crude oil. Spot and futures markets also exist for these products. An interesting feature of products derived from crude oil is that their price is dependent on the price of crude oil and also on the supply and demand for the particular product. Indeed the difference between the price of crude oil and a specific derived product such as heating oil or gasoline is often referred to as the crack spread. Since the crack spread can have a significant impact on profitability especially of refiners, a market in crack spread options has emerged. For example, on NYMEX it is possible to trade crack spread options on heating oil and unleaded gasoline.

1.4.2 Natural gas

Natural gas accounts for almost a quarter of world's energy consumption. The USA, accounts for around 30% of total consumption, Europe approximately 20% and the Russian Federation 15%. Until the 1970s, European and North American governments considered natural gas as a strategic product and thus a natural monopoly with state owned companies running the industry. However, following the energy shortages and crisis precipitated by OPEC sharply increasing the cost of oil during the 1970s, natural gas and the energy sector in general underwent structural reform to open the markets to competition in order to cut costs and improve economic performance and efficiency. Thus the main policy trend in natural gas markets over the recent decades has been the liberalization of the market. In both Europe and North American the natural gas sector has now fully evolved from an almost totally regulated industry to one that today operates largely as a free market. In the USA the process towards a more competitive market began with the enactment of the Natural Gas Policy Act of 1978. The Act, targeted at raising wellhead[5] prices, specified gradual price increases for various categories of gas, based on the expected price of oil.

Natural gas is traded on spot, futures, and options markets. The Henry Hub is the largest centralized point for natural gas spot and futures trading

in the USA. NYMEX launched the world's first natural gas futures contract in the early 1990s, it uses the Henry Hub as the point of delivery. NYMEX options on natural gas futures soon followed. Today the NYMEX gas futures contract trades 72 months into the future alongside a wide variety of other natural gas futures and options contracts traded on both NYMEX and the IPE. As with crude oil the price for natural gas fluctuates based on the changes in either supply or demand.

1.4.3   Electricity

Electricity, unlike crude oil and natural gas, cannot be stored and therefore must be consumed as it is produced. It can only be delivered to customers via wires, which also imply a limit on capacity. If these wires are at full capacity, it is possible that additional power is available: however, there is no means to transmit it. As a consequence of the continuing liberalization in energy markets futures and forward contracts have emerged for electricity. The Nordic Power Exchange or Nord Pool was the world's first international commodity Exchange for electrical power. As with any exchange it provides standardized contracts for both physical delivery and cash delivery. It is focused on providing services to Nordic nations (Sweden, Finland, Norway, and Denmark). NYMEX also offers a futures contract on US electricity (PJM Electricity Futures) along with a number of option contracts.

1.5   THE SCIENCE OF ENERGY RISK MODELING

Statistical methods provide the quantitative technology for empirical science by offering the energy risk manager the logic and methodology for the measurement of risk and for an examination of the consequences of that risk on the day to day activity of the business. In recent years it has become increasingly evident that knowledge of statistical methods is now an important job requirement. A responsible institution will need to assess its level of risk and monitor its level of price risk. Energy price risk management often involves a number of stages:

1. Analysis of energy risk via energy price risk modeling

2. Development of projected budgets and potential exposure

3. Identification of risk mitigation options – typically this might include hedging exposure by purchasing forwards, futures, or options in order to gain budget stability and predictability.

In this text we focus primarily on the first stage, energy price risk modeling with the objective of introducing many of the important statistical concepts and procedures required in order to begin the process of accessing and controlling energy price risk.

The evolving energy marketplace has become increasingly complex and dynamic. Unleaded gasoline, electricity, natural gas, heating oil and crude oil are exchange-traded commodities subject to frequent price swings on a short- and long-term basis. This volatility creates significant speculative opportunity and risk. Whilst it may be obvious that extreme levels of price volatility increase energy price risk, the risk to an institution's financial stability resulting from adverse movements in the level of energy prices, they also amplify other types of risk. For example, electricity price spikes in the Mid West of the USA during the summer of 1998 resulted in a sudden increase in energy price risk. The price spike was so large that a number of energy players found themselves exposed as a result of their writing (selling) uncovered options. Therefore the possibility of a default on outstanding obligations (i.e. credit risk) increased. In this instance we see clearly how increases in the level of one type of risk, can have knock-on effects on other types of risk. As a further illustration, operational risk events such as power plant outages, often simultaneously increase the likelihood of outages in other power plants (thus increasing operational risk) and lead to price spikes (thus increasing energy price risk), which in turn can lead to some energy players being exposed because of their derivative liabilities (and thus credit risk is increased). Therefore, whilst we focus in this text primarily on price risk modeling, it is important to remember that energy risks are often interrelated; when a risk event occurs it can have a knock-on effect on other types of risk.

1.6 FURTHER READING AND RESOURCES

The Union of Concerned Scientists (www.ucsusa.org) offers a balanced and informative historical account of the development of energy. They also provide detailed analysis of the current economic and scientific debate surrounding aspects of energy consumption. Three useful sites for information on the economics of crude oil are:

- ■ The American Petroleum Institute – www.api-ec.api.org

- ■ OPEC – www.opec.org

- ■ Independent Petroleum Association of America – www.ipaa.org

Further information about policy and economics of natural gas market can be found at www.oecd.org. Also take a look at www.ferc.gov and http://europa.eu.int. Web links to some of the key energy exchanges are given below:

- ■ IPE – www.theipe.com

- NYMEX – www.nymex.com

- The Tokyo Commodity Exchange – www.tocom.or.jp

- Nord Pool – www.nordpool.no

- Singapore exchange – www.sgx.com

Additional background on the history and development of energy can be found in Rudolph and Ridley (1986), Yergin (1992), Bower (1994), and Smil (1994).

Brower, Michael (1994) *Cool Energy: Renewable Solutions to Environmental Problems*. Massachusetts, MIT Press.
Smil, Vaclav (1994) *Energy in World History*. Westview Press, Colorado.
Yergin, Daniel (1992) *The Prize: The Epic Quest for Oil, Money, and Power*. Touchstone, New York.
Rudolph, Richard and Ridley, Scott (1986) *Power Struggle: The Hundred-Year War over Electricity*. Harper and Rowe, New York.

This page intentionally left blank

Statistical Foundations of Energy Risk Modeling

Around 100 years ago H.G. Wells noted that "statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." He made no mention of financial risk management because the financial market place was still in its infancy. Were he to make the comments today about risk management he would probably say that "statistical thinking is necessary not only for efficient citizenship, but also for effective financial risk management." It is now widely acknowledged that the effectiveness of energy risk management depends crucially on the soundness of the modeling used to assess, monitor and control price risk. At the core of modern energy risk management lie a broad body of tools that are under-pinned by statistical theory. In Part 1 of this book we introduce some of these tools, which together form the statistical foundations of energy risk modeling – applied probability (Chapter 2), descriptive statistics (Chapter 3) and inferential statistics (Chapter 4).

This page intentionally left blank

Introduction to Applied Probability for Energy Risk Management

There are many instances where those involved in energy products must make decisions under conditions of uncertainty. An oil producer must decide how much inventory to stock; a risk manger how much economic capital to set aside, and an electricity speculator when to buy or sell. In each of these cases the individuals make their decision on the basis of what they think is likely to occur; their decision is based on the probability that certain events will or will not happen. Most of us have some intuitive understanding of probability. Some people prefer to take the train to their place of work in the knowledge that a serious accident is less likely than if they drive. Others participate in high risk sports such as boxing or sailing, knowing that they are likely to face serious injury or death, but then again the likelihood of such extreme outcomes is actually quite small. Millions of individuals purchase lottery tickets even though the likelihood of wining a very large pay-out is extremely small. If we say that the probability of snow today is one-half, but tomorrow it is only one quarter, we know that snow is more likely today than tomorrow.

2.1 DESCRIBING RANDOM EVENTS

Underling the idea of probability is the notion of a random variable and random experiment. A random variable is a variable which can take on a given set of values. The spot price of natural gas, 3-month forward price

of crude oil, and the spread between the price on the nearest heating oil futures contract and the nearest Henry Hub Natural Gas futures contract are all examples of random variables. A random experiment, also known as an experiment of chance, is the process by which a specific value of a random variable arises. A random experiment is often seen as one of a sequence of such experiments known as a trial. The everyday battle between supply and demand in financial markets can be seen as a random experiment that yields outcomes (prices) for specific financial contracts. The application over a sequence of days, weeks, months, or years can be seen as a trial.

The outcome or collection of outcomes of a random experiment is called an event, with the simple event being an outcome that cannot be broken down further. The events of a random experiment are part of a larger sample space, which in turn consists of all possible outcomes for an experiment. In this sense the sample space consists of all possible simple events. For the most part we shall denote a random variable by a capital letter usually X and the value it takes by a lower case letter, that is, x. When a random variable X is observed on N occasions we obtain a succession of values denoted by $\{x_1, x_2, x_3, ..., x_N\}$; each of which provides us with a realization of X at specific points in time. The following example illustrates these concepts.

Example 2.1   Price of nearest NYMEX Henry Hub Natural Gas futures contract as a random variable

During the trading day the price of the nearest NYMEX Henry Hub Natural Gas futures contract varies from one hour to the next. In the language of statistics the price at any moment of time is a random variable whose sample space consists of all possible settlement prices. The trading process is the experiment, the passage of time from one moment to the next a trial, the settlement price at the end of the day the outcome. At the start of day the experiment (trading) begins. At this stage the outcome (settlement price) of the experiment is unknown. Will the settlement price be higher, lower, or remain the same by the end of trading? At the end of the trading day the outcome of the experiment, the observed settlement price, is known. The simple event in this case is the settlement price on a particular day. Suppose we observed the settlement price of the nearest contract for five days between June 10th and 14th as {$5.162, $5.151, $5.150, $5.155, $5.433}. We may also write this sequence as $\{x_1 = \$5.162, x_2 = \$5.151, x_3 = \$5.150, x_4 = \$5.155, x_5 = \$5.433\}$. From this simple event we can construct more complicated events. For example, the average settlement price over the past five trading days or the difference between the settlement price on Mondays and Fridays. We now see that a random variable describes the value that corresponds to the outcome from a given experiment. A random variable is actually a function that associates a unique numerical value to

every outcome of a random experiment. The term random is used to remind us that we will not know what the value of the random variable is until after the experiment has been conducted.

There are two basic types of random variable – discrete and continuous. A discrete random variable may take on a countable number of distinct values. A continuous random variable is one that can take on any real value. In Example 2.1 the settlement price of the nearest NYMEX Henry Hub Natural Gas futures contract is a continuous random variable. Suppose we are interested in whether or not the price was greater than $5.154; is this also a continuous random variable? No because it can only take on the states *yes* or *no.* To see that this is the case recall the settlement price of the nearest NYMEX Henry Hub Natural Gas futures contract between June 10th and 14th is a continuous random variable with values $\{x_1 = \$5.162, x_2 = \$5.151, x_3 = \$5.150, x_4 = \$5.155, x_5 = \$5.433\}$. For the same days, the random variable *"greater than $5.154"*, would take the states {yes, no, no, yes, yes}. If we denote "yes" by 1 and "no" by 0, we can rewrite this as {1,0,0,1,1}. As this random variable can only take on two values it is discrete.

2.2   WHAT IS PROBABILITY?

The only time we need to make a decision is when we are uncertain about a situation. Consider the scenario where your boss asks you to assess the risk in a new financial product and report back to him with a recommendation by the following day. You begin your research, but find nothing like it in the current market place or in the academic or professional literature. You have reached the proverbial "fork in the road." A decision has to be made. Do you ask for more time, reject the product, or cross your fingers and accept it? Either way a decision has to be made because you do not know something. That is, you are uncertain about something related to the new product. Probability offers a formal structure for describing uncertainty.

2.2.1   Properties of probabilities

The development of a mathematical theory of probability began during the seventeenth century with the French nobleman Antoine Gombauld, who puzzled over the likelihood of obtaining two 6s at least once in 24 throws of a pair of dice. We have seen that random variables can be used to describe the possible values of future events. Probability attaches a likelihood to each of the possible values. For example we may not know for sure whether it is going to snow today, but after hearing the weather forecast and looking out of the window, we might conclude the probability of snow is 6/10 or 60%.

In this case the random variable is "snow today?" which can take the states "yes" and "no." To the state "yes" we have attached a probability of 60%. Intuitively, an estimate of probability should lie between 0 and 1. This implies a probability of 40% to the state "no." Thus the probability of snow today plus the probability of no snow today is equal to 1.

Probability offers a formal mechanism for quantifying and expressing uncertainty inherent in decision making. To illustrate this further let us suppose that we are interested in whether or not today's settlement price of the nearest Henry Hub Natural Gas futures contract is greater than $5.154; we shall denote this event by H and use Prob(H) as shorthand for the probability of observing H. What is the probability of not observing H (i.e. the price being less than or equal to 5.154)? This is the complementary event and we write its probability as Prob($\sim H$). Since the events H and $\sim H$ are mutually exclusive, that is if H occurs $\sim H$ cannot occur, we can write:

- Prob(H) + Prob($\sim H$) = 1 – which informs us that the sum of the probabilities of the mutually exclusive events H and $\sim H$ should sum to one. In other words the probability that the settlement price is less than or equal to 5.154 or it is greater than 5.154 is 1.

- $0 \leq$ Prob(H) ≤ 1 which satisfies our instinct that the probability of an event always lies between zero and one.

An event that is impossible has a probability of 0, and an event which is certain to occur will have a probability of 1. In this sense probability indicates the relative likelihood of an event happening. The closer the probability is to 1 the more likely the event is to occur, and the closer to 0, the less likely the event is to occur. What is the probability that today's settlement price of the nearest NYMEX Henry Hub Natural Gas futures contract will be the same as yesterday, equal to yesterday, or more than yesterday? Since one of these outcomes is certain to occur the probability is 1. In summary, probability provides a well-defined numerical scale for measuring uncertainty inherent in the outcome of random variables. We will make use of it to help characterize the risk inherent in energy spot and derivative prices. What is the probability that the Henry Hub Natural Gas futures contract is greater than $5.154? To answer this question we will need to address the question of where we get probabilities from.

2.2.2   Distinction between empirical and subjective probabilities

Values for the probabilities of events of interest can be derived from empirical data or elicited from expert opinion. Probability values derived from

empirical observations reflect the relative frequency of the occurrence of an event after a very large number of random trials – it is the proportion of times an event can be expected to occur. As an example, consider a risk manager who counts the number of times the price of the nearest IPE Gas Oil futures prices moved by more than 15% in a week. Suppose the risk manager collects data over a period of 100 weeks, of which 5 weeks contain a move in price of 15% or more, then the empirical probability of a 15% or more price change can be calculated as: $5/100 = 0.05$ or 5%, empirically derived probabilities give an objective status to the notion of probability by rendering it a property of real world phenomena. In this case probability is interpreted as stemming from the observable stability of empirical frequencies. These empirical frequencies can be approximated using mathematical probability models; we discuss this issue further in Section 2.3 and Chapter 5.

Probabilities elicited from experts are known as subjective probabilities. A crude oil speculator might conclude that the probability of a significant price increase in the spot price of oil is 95%, following the announcement of a series of oil refinery explosions in the Middle East. Another trader with the same information might conclude it is only 65%. Subjective probability renders the notion of probability a subjective status by regarding it as "degrees of belief" on behalf of individuals assessing the uncertainty of a particular situation. Its value depends heavily on personal viewpoints and prior experience.

2.2.3   Some rules for the computation of probabilities

Many computations involving probabilities can be achieved using five basic rules:

- Rule 1 – Addition rule for mutually exclusive events
- Rule 2 – Addition rule for dependent events
- Rule 3 – Multiplication rule for statistically independent events
- Rule 4 – Multiplication rule for dependent events
- Rule 5 – Conditional probability.

Rule 1, the addition rule for mutually exclusive events, has to do with the probability that one or the other of the events will occur. Two events are mutually exclusive if the occurrence of one event precludes the occurrence of the other event. The rule tells us that if two events A and B are mutually exclusive, the probability that one or other of the events will occur is the

sum of their separate probabilities:

$$\text{Prob}(A \text{ or } B) = \text{Prob}(A) + \text{Prob}(B)$$

For three mutually exclusive events, $A, B,$ and $C,$ the rule is written:

$$\text{Prob}(A \text{ or } B \text{ or } C) = \text{Prob}(A) + \text{Prob}(B) + \text{Prob}(C)$$

**Example 2.2   *Mutually exclusive events and the nearest
NYMEX PJM Electricity Futures contract***

Suppose the February NYMEX PJM Electricity Futures contract is currently trading at \$45.55. A speculator might be interested in whether the settlement price in two days time is higher, lower, or equal to the current price. Let us label these events as A, B, and C respectively. Let us suppose further, that over the past 400 trading days the speculator observes the values shown in Table 2.1.

What is the probability that the settlement price will be less than or greater than the current price? The outcome "less than \$45.55" is event A. The outcome "greater than \$45.55" is event C. Applying the addition rule for mutually exclusive events and using the empirically derived probabilities we find

$$
\begin{aligned}
\text{Prob}(A \text{ or } C) &= \text{Prob}(A) + \text{Prob}(C) \\
&= 0.025 + 0.075 \\
&= 0.10 \text{ or } 10\%.
\end{aligned}
$$

We would also expect

$$\text{Prob}(A) + \text{Prob}(B) + \text{Prob}(C) = 1,$$

Table 2.1 NYMEX PJM Electricity futures price observations

Price ($)	Event	Number of days event observed	Probability of occurrence
Less than 45.55	A	10	10/400=0.025
Equal to 45.55	B	360	360/400=0.90
Greater than 45.55	C	30	30/400=0.075

and this is indeed the case because either A, B, or C is certain to occur. Thus we see

$$\text{Prob}(A) + \text{Prob}(B) + \text{Prob}(C) = 0.025 + 0.90 + 0.075 = 1.$$

In addition given $\text{Prob}(A)$ and $\text{Prob}(C)$ we can find $\text{Prob}(B)$ which is equal to

$$1 - [\text{Prob}(A) + \text{Prob}(B)]$$
$$= 1 - [0.025 + 0.075]$$
$$= 1 - 0.1$$
$$= 0.9 \text{ or } 90\%.$$

Rule 2, the addition rule for dependent events, informs us how to calculate probabilities when two events, A and B, are not mutually exclusive – that is, when the probability of occurrence of one event is dependent or conditional on the occurrence or nonoccurrence of the other event. In this case it is possible for both events to occur. The probability of both A and B occurring is known as the joint probability and denoted by $\text{Prob}(A \text{ and } B)$. In this situation the probability that one or other of the events will occur is

$$\text{Prob}(A \text{ or } B) = \text{Prob}(A) + \text{Prob}(B) - \text{Prob}(A \text{ and } B)$$

Notice if A and B were mutually exclusive

$$\text{Prob}(A \text{ and } B) = 0$$

and hence we obtain rule 1.

Rule 3 makes use of the notion of statistical independence. Two events are said to be statistically independent if the occurrence or nonoccurrence of one event does not affect the probability that the other event will occur. If two events A and B are statistically independent then the probability that they will both occur is the product of their separate probabilities

$$\text{Prob}(A \text{ and } B) = \text{Prob}(A) \times \text{Prob}(B)$$

This rule for combining probabilities presumes that the outcome of A is not affected by the outcome of B. To further illustrate what is meant by independence of outcomes, suppose we consider the spot price of natural gas and the spot price of the dollar–sterling exchange rate at the close of business on any particularly day. There is no reason to suppose that the level of the exchange rate and the spot price of natural gas are in any way related. Hence, the price (outcome) at the end of the trading day of natural gas is unaffected by the level (outcome) of the dollar–sterling exchange rate.

To put it another way, the two events are independent because the outcome of the first event (spot price of natural gas) does not depend on the outcome of the second event (level of the exchange rate).

Rule 4, the multiplication rule for dependent events, informs us how to calculate joint probabilities when events are dependent. If two events A and B are dependent, the probability they will both occur is

$$\text{Prob}(A \text{ and } B) = \text{Prob}(A|B) \times \text{Prob}(B)$$

Notice that we use the notation $\text{Prob}(A|B)$ to denote the probability that A will occur given that B has already occurred. The vertical line means "given that."

We can rewrite Rule 4, to give us Rule 5, the law of conditional probability

$$\text{Prob}(A|B) = \frac{\text{Prob}(A \text{ and } B)}{\text{Prob}(B)}$$

The rule tells us that the conditional probability of an event A given another event B is the probability that A will occur given that B has occurred.

Example 2.3   Conditional probability of the spot price of electricity given the Central Appalachian May Coal Futures settlement price

Consider the risk manager of a coal mining company that also operates electric power plants. Suppose the manager is interested in calculating the conditional probability of the spot price of electricity being below $48 given the future price of coal is below $45.35. More specifically denote the event *May Central Appalachian Coal Futures price is less than $45.35* as B and the event *May Central Appalachian Coal Futures price is less than $45.35 and spot electricity price* is below $48 as A. Let the probability that the May futures coal price falls below $45.35 be 20%, and 5% for the probability that May futures coal price falls below $45.35 and the spot electricity price falls below $48. So,

$$\text{Prob}(B) = 20\%$$

and

$$\text{Prob}(A \text{ and } B) = 5\%.$$

Thus the probability that the electricity spot price falls below \$48 given that we observe the May futures coal price is below \$45.35 is

$$\text{Prob}(A|B) = \frac{\text{Prob}(A \text{ and } B)}{\text{Prob}(B)} = \frac{0.05}{0.2} = 0.25.$$

If we rearrange the above formula we see

$$\text{Prob}(A \text{ and } B) = \text{Prob}(B)P(A|B)$$

and so,

$$\text{Prob}(A \text{ and } B) = \text{Prob}(B)\text{Prob}(A|B) = 0.2 \times 0.25 = 0.05,$$

which is as expected.

2.3   PROBABILITY FUNCTIONS

One of the most common mechanisms for expressing uncertainty is via a probability function. Formally, probability functions are mathematical functions used to characterize all the uncertainty surrounding the outcome of a random variable. But what exactly does this mean? Well, the uncertainty surrounding an unknown event can be summarized in an appropriate mathematical function. Such functions inform us what outcomes are possible for a given random variable and their associated probability. This is important because once we have knowledge of the potential outcomes of an event (random variable) and their associated likelihood we can begin to quantify energy price risk. We shall consider three broad types of probability function:

1. Probability distributions
2. Cumulative probability functions
3. Inverse probability functions.

The probability of a random variable taking a particular value is captured in a probability distribution. A probability distribution is a mathematical model that maps all of the outcomes that a random variable can take to specific probabilities. It thereby gives us a complete listing of all possible outcomes of an experiment together with their probabilities. In this sense probability distributions provide a complete representation of the

uncertainty surrounding every potential outcome of a random variable – it informs us what outcomes are possible and how likely they are. There are two broad types of probability distribution, those for discrete random variables known as probability mass functions and those for continuous random variables know as probability density functions.

2.3.1   Probability mass function

A probability mass function yields the probability with which a discrete random variable can take a particular value. For example, we previously saw that the event *whether or not the price of the nearest NYMEX Henry Hub Natural Gas futures contract is greater than $5.154* is a discrete random variable which can only take on two values *yes* or *no*. We denote this event by H, which takes the value 1 if the event occurs and 0 otherwise. The complete probability mass function $p(x)$ takes the numerical values

$$\text{Prob}(X = 0) = p(0) = \frac{9}{10}$$

and

$$\text{Prob}(X = 1) = p(1) = \frac{1}{10}$$

In this case $p(X = 0)$ and $p(X = 1)$ make up the probability distribution for the random variable. Figure 2.1 illustrates the probability mass function for this example. The low value of $p(X = 1)$ informs us that the event occurs with low probability and is in this sense unlikely. Notice we have assumed that we know the values of $p(X = 0)$ and $p(X = 1)$; in practice we will need to estimate these values from empirical data or our own personal beliefs. If the probability function is based on personal beliefs it is likely to have different values, say for a trader who is hoping for a large increase in price, and a hedger who is hoping for price stability. It should also be noted that the total of the probabilities sum to 1. This will be true for all probability distributions. This is because the distribution is a complete listing of all outcomes of the experiment of chance – the probability that one or other of these outcomes will occur is a certainty. There are many probability distributions that are important in understanding and applying statistical methods. Binomial, Poisson Geometric, Hypergeometric, Pascal, Polya–Aeplli, Poisson Inverse Gaussian, Hofmann, Neymann Type A, and Neymann Type A plus Poisson are just some of a very large number of discrete probability distributions.
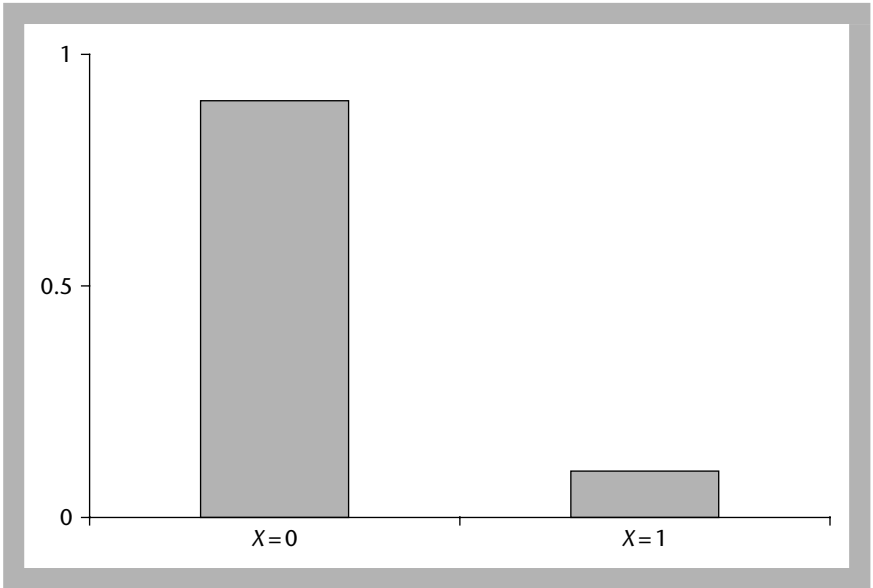
Figure 2.1 Discrete probability distribution for the simple event of whether or not the price of Henry Hub is greater than $5.154

2.3.2 Probability density function

A probability density function yields the probability with which a continuous random variable lies within a specified range. Given a random variable X, we denote its associated probability density function by $f(x)$. Figure 2.2 illustrates a probability distribution for a continuous random variable. It represents the scattering of likely values along the horizontal axis. That is the horizontal axis represents the range of plausible outcomes for the random variable of interest. The vertical axis provides a measurement of the likelihood of each outcome. Continuous probability distributions include the Normal, Lognormal and Beta distributions.

Since $f(x)$ is a continuous function the area between any two points say x_1 and x_2 represents the probability that the random variable will lie between these two values. We write this as

$$\text{Prob}(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f(x)dx$$

Probabilities are equal or greater than zero; thus the probability mass function and probability density function satisfy

$$p(x) \geq 0 \quad \text{and} \quad f(x) \geq 0$$
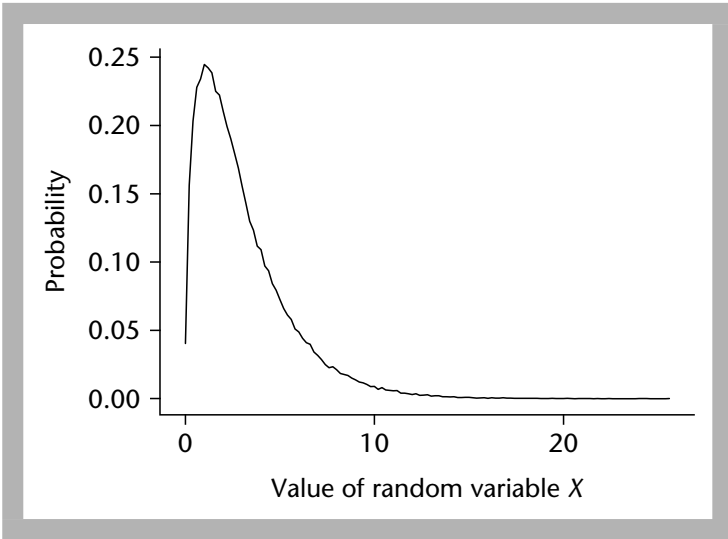
Figure 2.2 Probability distribution of a continuous random variable

In addition, we have already seen that the sum over all possible outcomes is equal to 1 so that

$$\sum_x p(x) = 1$$

and

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

2.3.3 Cumulative distribution function

For a random variable X the cumulative distribution function measures the cumulative probability and is calculated by

$$F(X \leq x) = \text{Prob}(X \leq x) = \begin{cases} \sum_{x \leq k} p(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{x} f(u)du & \text{if } X \text{ is continuous} \end{cases}$$

If the random variable X is discrete the cumulative distribution function $F(x)$ is a step function as illustrated in Figure 2.3. If X is continuous then $F(x)$
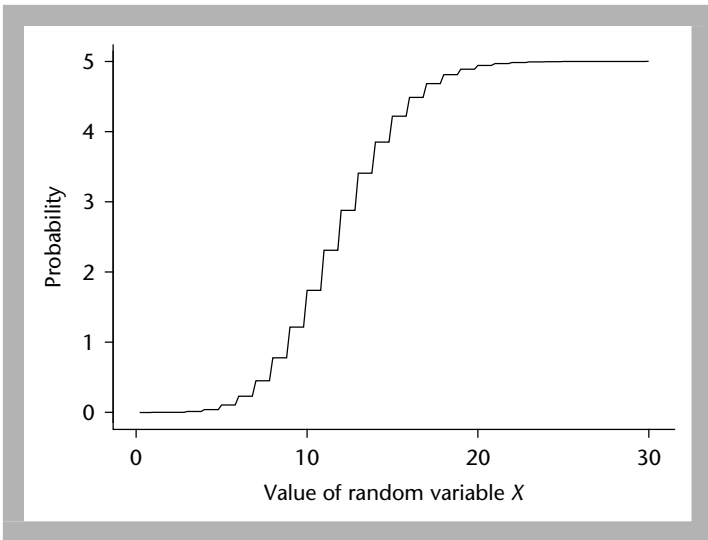
Figure 2.3 Cumulative distribution of a discrete random variable
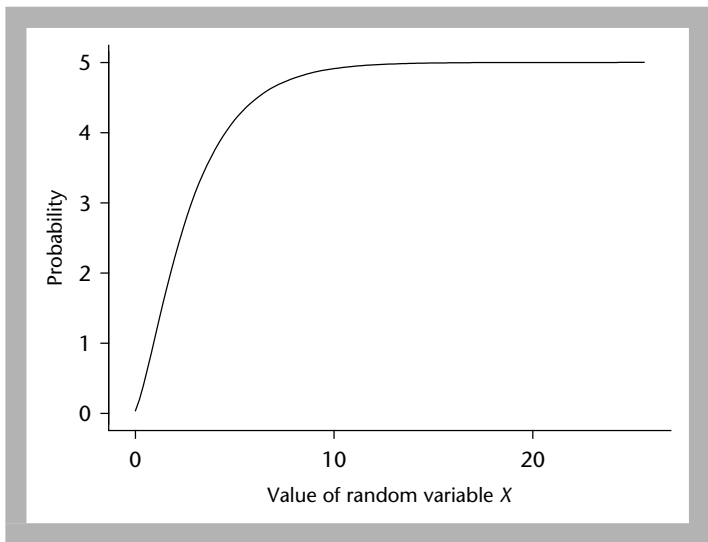
Figure 2.4 Cumulative distribution of a continuous random variable

is a continuous function as illustrated in Figure 2.4 It is also worth noting that given a probability density function $f(x)$ the cumulative function can be defined as

$$F(x) = \int_{-\infty}^{x} f(u)du.$$

2.3.4   Inverse probability function

Given a random variable X the αth percentile ($0 \leq \alpha \leq 1$) is that value of the random variable X say x_α, that indicates the percent of observations that is equal to or below x_α. Given the distribution function $F(x)$ the percentile function, also known as the inverse probability function, is denoted by $F_x^{-1}(1 - \alpha)$. To illustrate this concept, suppose X is a random variable of the daily high spot price for Brent Crude Oil. Over the past 100 days we will record $\{x_1, ..., x_{100}\}$, where x_1 is the observed high on the first day and x_{100} the number observed on the 100th day. If p represents the observations arranged into ascending order, so that p_1 is the smallest and p_{100} the largest, the 99th percentile is equal to or greater than 99% of the values recorded in $\{p_1, ..., p_{100}\}$. This is p_{99}.

2.4   THE NORMAL DISTRIBUTION

The question that now arises is what is a suitable probability distribution for the random variable X? When mathematicians of the eighteenth century began to address this question the Normal distribution emerged.[1] A continuous random variable, X, is said to follow a normal distribution with mean, μ, and standard deviation, σ, if it has the probability density function given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right), \quad -\infty < x < \infty.$$

The formula for the normal distribution actually defines a family of distributions parameterized by μ and σ. As illustrated in Figure 2.5 the normal distribution is symmetric with observations more concentrated in the middle of the distribution than the tails. The parameter μ is equal to the arithmetic mean and is known as a location parameter because changing the mean shifts the curve along the x-axis. The parameter σ is the standard deviation and is frequently termed a scale parameter because changing the standard deviation changes the spread of the curve.

The most flexible probability distributions will have at least one parameter for location, scale and shape. The location parameter controls where on the horizontal axis the distribution is centered, the scale parameter controls the spread of probability around the center of the distribution and the shape parameter controls the shape of the distribution. The normal distribution only has parameters for location and scale. Despite this it plays a central role in statistical theory and practice, particularly in the area of inferential statistics.
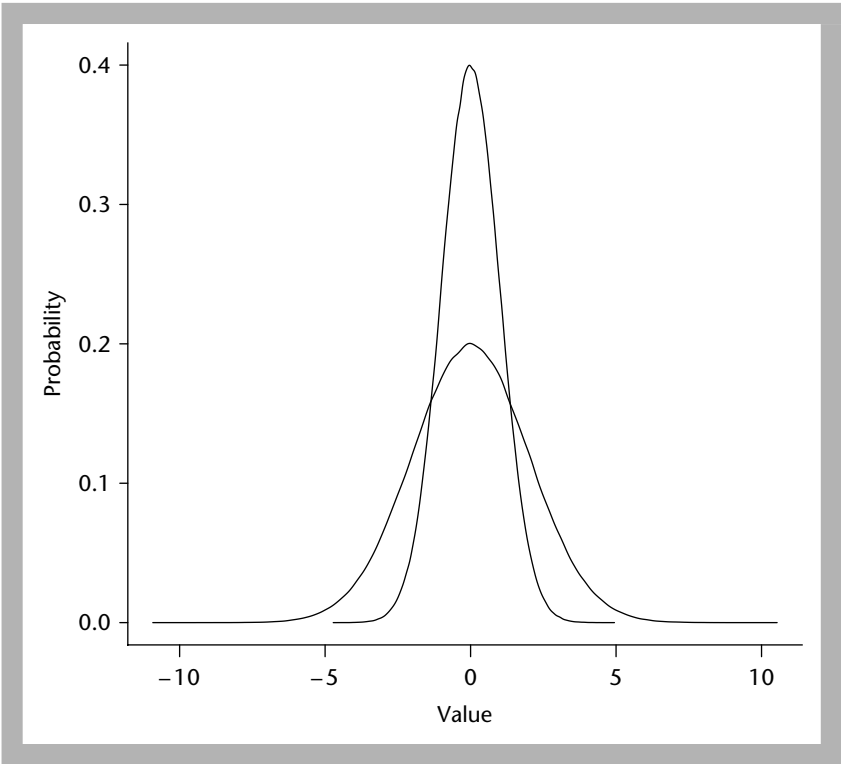
Figure 2.5 Two normally distributed random variables with the same mean but different levels of dispersion

2.4.1   Central limit theorem

Given a sequence of independently identically distributed random variables $\{X_1, ..., X_N\}$ with mean μ and variance σ^2 the central limit theorem informs us that the distribution of the arithmetic mean,

$$\overline{X} = \frac{X_1 + \cdots + X_N}{N},$$

is approximately normally distributed as N becomes increasingly large. The central limit theorem therefore informs us what the distribution of $\overline{X}$ the sample mean is. Provided N is large it will be approximately normally distributed almost regardless of distribution of the individual observations $\{X_1, ..., X_N\}$.

2.4.2   Transformation to the standard normal distribution

If X is normally distributed with mean μ and variance σ^2 it can be transformed into a standard normal random variable Z via

$$Z = \frac{X - \mu}{\sigma}.$$

The transformed random variable Z, also known as a standardized variable or standard normal variable has a mean of zero and a variance equal to 1. Standardized values are sometimes called Z scores. What is the value of using a Z score? Z scores reflect the number of standard deviations above or below the mean that a particular value lies. For instance a Z score of -0.62 indicates that this particular observation is -0.62 standard deviations below the mean. Some values of the cumulative probability function for the standard normal distribution are given below

z-score	F(z)
−3.0	0.0013
−2.5	0.0062
−2.0	0.0228
−1.5	0.0668
−1.0	0.1587
−0.5	0.3085
0.0	0.5000
0.5	0.6915
1.0	0.8413
1.5	0.9332
2.0	0.9772
2.5	0.9938
3.0	0.9987

Thus a Z score of -3 implies $F(z) = 0.0013$, whilst a value of 0 implies $F(z) = 0.5$.

2.5   RELEVANCE OF PROBABILITY FOR ENERGY RISK MANAGEMENT

Managing energy risk involves assessing and controlling price risk. Price risk arises because the future price of an energy product, such as coal or crude oil is uncertain. Through the concepts of random variables and probability discussed so far we see that the price of a energy forward, future, or spot contract is a random variable whose likely future price can be measured

using probability derived from an appropriate probability model. Probability distributions provide a compact notation for describing the likelihood of all of the potential future outcomes. In this sense they provide a complete characterization of the uncertainty surrounding future changes in energy prices. Price risk is caused by movements in the prices of energy products. In discussing price risk it is often more convenient to work with price returns rather than price levels. We define the simple return as

$$\tilde{r}_{t+1} = (S_{t+1} - S_t)/S_t,$$

where t represents time and S is the price of the energy product. Hence S_t could be regarded as the price today and S_{t+1} the price tomorrow. It is often more convenient to use the log return defined by

$$r_{t+1} = \ln(S_{t+1}/S_t),$$

where $\ln()$ is the natural logarithm.

When the difference between t and $t+1$ is small, $\tilde{r}_{t+1}$ and r_{t+1} are fairly similar because

$$\tilde{r}_{t+1} = (S_{t+1} - S_t)/S_t = S_{t+1}/S_t - 1,$$

and

$$\ln(S_{t+1}/S_t) \approx S_{t+1}/S_t - 1 \quad \text{provided } S_{t+1}/S_t \text{ is close to 1.}$$

The advantage of using the log return is that we can calculate the compounded return at day M as the sum of the daily returns. Table 2.2 illustrates this point. It shows the daily 2-month forward Brent Crude Oil price alongside the simply and log return. Using the simple return, we can see that the price has risen by:

$$\frac{24.35 - 20.16}{20.16} = 20.78\%.$$

However, the sum of the simple arithmetic returns is $0.30\% + 0.49\% + \cdots + (-1.77\%) = 19.26\%$, which is somewhat less than the actual simple return. Using the log return, we see that over the period the price has risen by

$$\frac{\ln(24.35)}{\ln(20.16)} = 18.88\%,$$

and the sum of the log returns is $0.30\% + 0.49\% + \cdots + (-1.79\%) = 18.88\%$

Table 2.2 Simple and log return of the 2-month forward price of Brent Crude

Date	2-month forward Brent Crude Price (FOB U$/BBL)	Simple return (%)	Log return (%)
02/20/2002	20.16		
02/21/2002	20.22	0.30	0.30
02/22/2002	20.32	0.49	0.49
02/25/2002	20.14	−0.89	−0.89
02/26/2002	21.07	4.62	4.51
02/27/2002	20.93	−0.66	−0.67
02/28/2002	21.40	2.25	2.22
03/01/2002	22.02	2.90	2.86
03/04/2002	22.18	0.73	0.72
03/05/2002	23.08	4.06	3.98
03/06/2002	23.23	0.65	0.65
03/07/2002	23.58	1.51	1.50
03/08/2002	23.70	0.51	0.51
03/11/2002	24.28	2.45	2.42
03/12/2002	24.30	0.08	0.08
03/13/2002	23.94	−1.48	−1.49
03/14/2002	24.38	1.84	1.82
03/15/2002	24.18	−0.82	−0.82
03/18/2002	24.78	2.48	2.45
03/19/2002	24.79	0.04	0.04
03/20/2002	24.35	−1.77	−1.79

2.6   A PROBABILISTIC MODEL FOR ENERGY PRICE RISK

Having laid the probabilistic foundations, and given basic definitions of the log return and simple return we are now in a position to begin our formal analysis of energy price risk. What we require is a model that will capture the price risk inherent in an energy contract. As we have seen probability provides the required tools to develop such a model. We begin with the notion of a random variable R representing the uncertain future return of an energy product of interest such as the coal spot price, prompt month crude

futures price, natural gas forward price etc. The return may be calculated on a daily, weekly, monthly, or other basis. For now we shall assume the random variable R is continuous with probability density function, probability function and inverse probability function given by $f(r)$, $F(r)$ and $F_R^{-1}(1 - \alpha)$ respectively. As we have seen these functions describe in its entirety the uncertainty surrounding R. Our probabilistic model of R thus takes the generic form

$$R \sim f(r)$$

where $\sim$ means "distributed" and indicates that R is from some probability function. If we assume R is from the normal distribution, with parameters $\mu = 0$ and $\sigma = 1$ we can write

$$f(r) = N(\mu = 0, \sigma = 1),$$

or more compactly

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

This provides us with our first model of energy price risk. Our next objective will be to assess the validity and where necessary improve on this basic model. As a first step in order to achieve this we turn in Chapter 3 to the field of descriptive statistics.

2.7   SUMMARY

Random variables and probability provide us with the basic probabilistic tools required to gain insights into the nature of the uncertainty surrounding energy price risk. We have seen that it is possible to capture such uncertainty via probability. Probabilities may be classified as empirical and subjective. Empirical probabilities are determined from observation and experimentation. Subjective probabilities are based on degrees of belief. We can capture the uncertainty surrounding future possible outcomes with a probability function. A probability function is a mathematical description of the likely outcomes of a random variable. However, probability is not the only tool required for effective modeling; knowledge of random variables and descriptive statistics is also important.

2.8   FURTHER READING

An applied and accesible introduction to probability theory can be found in Hines and Montgomery (1980). Practical applications relevant to risk management are given in Lewis (2003). Lewis (2004) provides hands on examples which can run in Microsoft Excel.

Hines, W. W. and Montgomery, D. C. (1980) *Probability and Statistics in Engineering and Management Science*. John Wiley and Sons, New York.
Lewis, Nigel Da Costa (2003) *Market Risk Modeling: Applied Statistical Methods for Practitioners*. Risk Books, London.
Lewis, Nigel Da Costa (2004) *Operational Risk with Excel and VBA: Applied Statistical Methods for Risk Management*. John Wiley & Sons, Inc., New York.

2.9   REVIEW QUESTIONS

1 Determine which of the following random variables are discrete and which are continuous:
  (a) The number of customers arriving at a check-out counter in an hour
  (b) The amount of oil exported from Iraq in a month
  (c) The difference between the level of inflation in the United Kingdom and United States of America
  (d) The number of fatal automobile accidents in Stamford, Connecticut in a given month
  (e) The volume of traded contracts on a futures exchange
  (f) The change in the spot price of natural gas in a day
  (g) The number of claims on a medical insurance policy in a particular week
  (h) The time that elapses between the installation of a new software package and its first failure.

2 List four examples of Random Experiments alongside their associated random variables. From your list identify the continuous and discrete random variables.

3 Explain the difference between subjective and empirical probability.

4 Describe each of the following and explain why they may be important tools for characterizing energy price risk:
  (a) Probability mass function
  (b) Probability density function
  (c) Cumulative distribution function
  (d) Percentile function.

5 Let $F( )$ be the cumulative Normal distribution function. Calculate:
  (a) $F(1.25)$
  (b) $F(-1.25)$
  (c) $F(2)$
  (d) $F(-2)$
  (e) $F(0)$
  (f) Comment on your findings.

Descriptive Statistics of Energy Prices and Returns

Descriptive statistics are those statistical methods that are used to summarize the characteristics of a sample. The main purpose of descriptive statistics is to reduce the original sample into a handful of more understandable metrics without distorting or losing too much of the valuable information contained in the individual observations. We begin by collecting N observations $\{r_1, r_2, ..., r_N\}$ on the price return random variable R. These measurements are then organized and summarized using techniques of descriptive statistics described in this chapter. In most cases we can compactly describe the characteristics of a sample using three distinctive classes of descriptive statistics. The first class summarizes the center of the distribution and are known as measures of central tendency. The second class summarizes the spread or dispersion of the sample and are commonly known as measures of dispersion. The third class known as shape statistics summarizes important elements of the shape of the underlying probability distribution implied by the sample. Our objective in using descriptive statistics is to describe as compactly as possible the key properties of empirical data. This information will then be used to assist us in selecting an appropriate probability model for modeling price risk.

3.1   MEASURES OF CENTRAL TENDENCY

Central tendency refers to where the bulk or typical observations in our sample of price return observations $\{r_1, r_2, ..., r_N\}$ lie. In everyday

speak central tendency refers to the "average" value in our sample. The primary measures of central tendency are the arithmetic mean and median.

3.1.1 Arithmetic mean

Given a sample of observations on the price return of an energy product $\{r_1, r_2, ..., r_N\}$ the arithmetic mean, which we denote by $\bar{r}$, is the sum of the set of observations in a sample divided by the number of observations

$$\bar{r} = \frac{r_1 + r_2 + \cdots + r_N}{N}$$

The formula for calculating $\bar{r}$ is known as an estimator, the actual value $\bar{r}$ takes is known as an estimate. Before moving on let us take a moment to reflect on the previous sentence. We have a sample of observations of price returns on a particular energy product. We wish to characterize the future uncertainty surrounding these price returns using probability. We also know that the appropriate probabilities can be obtained using the appropriate probability distribution. It turns out that a probability distribution is determined by a number of parameters. These parameters, often have a meaningful interpretation. For example, if we choose to model the price returns using a normal probability distribution, the arithmetic mean is one of the two parameters that determine the probability values of the distribution. The important point to take away from this discussion is that we often will use our sample to derive estimates of the parameters of some underlying probability model. In this chapter we also use sample estimates (that is estimates derived using an estimator and sample observations) to tell us something about the characteristics of the price return of the energy product in question.

Example 3.1   Arithmetic mean of the daily change spot price of natural gas

To illustrate the notion of an arithmetic mean, consider the monthly change in the spot price of natural gas over the five months from September to January $\{r_1 = 1\%, r_2 = 2\%, r_3 = 1\%, r_4 = 1\%, r_5 = 1\%\}$. Notice that r_1 is the return for the month of September and r_5 the return for January. The arithmetic mean return over the five-month period can be calculated as

$$\bar{r} = \frac{0.01 + 0.02 + 0.01 + 0.01 + 0.01}{5} = 0.012 \text{ or } 1.2\%.$$

3.1.2   Trimmed mean

The key advantage of using the arithmetic mean is that it is a widely used and understood measure. Its major weakness lies in its sensitivity to unusual observations (also known as outlier observations). To illustrate the sensitivity of this measure, suppose that due to a historically cold January $r_4 = 60\%$ rather than 1%. the arithmetic mean would then be calculated as

$$\bar{r} = \frac{0.01 + 0.02 + 0.01 + 0.60 + 0.01}{5} = 0.13.$$

Therefore the average return is 13%. Is this a good measure of the center of the sample? Seemingly not , because for most of the months the return was between 1% and 2%. The observation in January was unusually large – an outlier. One way to reduce the sensitivity of the arithmetic mean to extreme observations is to use a trimmed mean. A trimmed mean is calculated by discarding a given percentage of the lowest and the highest values in a sample and then computing the arithmetic mean of the remaining values. As such it is less susceptible to the effects of outliers than is the basic arithmetic mean. If we use a 20% trim on the above observations we have

$$\bar{r}_{\text{Trim}} = \frac{0.02 + 0.01 + 0.01}{3} = 0.0133.$$

Thus the trimmed mean is 1.33%. This proves to be a more satisfactory measure of the typical value in the sample.

3.1.3   Median

The key disadvantage of the trimmed mean is that it does not use all of the information in a sample. This is because it excludes observations from the calculation. An alternative, which uses all of the observations is the median. The median is the value that is greater than or equal to half of the values in the sample. For a continuous random variable with probability density function $f(x)$ it is the value such that

$$\int_{-\infty_1}^{x_{\text{median}}} f(x)\, dx = \int_{x_{\text{median}}}^{\infty} f(x)\, dx = \frac{1}{2}$$

As the median is the point between the lower and upper halves of a distribution it can be calculated by arranging the N observations of a sample in increasing order. The median is the $[(N+1)/2]$th observation when N is odd and the average of the Nth and the $[(N+1)]$th observation when N is even.

The median is known as a robust estimator because it is less sensitive to extreme observations than the mean. For this reason it is useful in situations where the sample includes outliers. Returning to the previous illustration, where we saw that an extremely cold January resulted in an extreme movement in the spot price of gas, the median is obtained by ranking the sample from lowest to highest

$$\{1\%, 1\%, 1\%, 2\%, 60\%\}.$$

Since N, the number of observations in our sample is equal to 5 the median is equal to the $(5 + 1)/2 = 3$rd observation. In this case the median return is equal to 1% which is slightly lower that the trimmed mean of 1.33 and much more representative of the sample values than the arithmetic mean of 13%.

3.2   MEASURES OF DISPERSION

Often observations in a sample of energy price returns will congregate closely together and in other cases they may be more dispersed, this is because energy prices vary over time with some time-periods more volatile than others. The degree of dispersion or variability leads to another important way in which our sample of price returns $\{r_1, r_2, ..., r_N\}$ can be characterized.

3.2.1   Simple range

Perhaps the simplest method of dispersion is obtained by calculating the difference between the minimum and maximum value in a sample, known as the range. Returning to Example 3.1, we see the range is

$$\text{Range} = 0.6 - 0.01 = 0.59.$$

The key advantage of the range is that it is easy to compute. However since it depends only on the largest and smallest observation it does not use all the information because it ignores observations in the middle of the sample. To see this suppose we have a sample of the spot price of crude over the same time-period with the price returns $\{1\%, 60\%, 59\%, 58\%, 60\%\}$. For crude oil the returns are much higher, clustering around 58%. The range is

$$\text{Range} = 0.6 - 0.01 = 0.59,$$

which is exactly the same value as for the spot price of natural gas. Yet the distribution of returns between the two products is very different.

3.2.2   Interquartile range

There are several ways to measure the variability of a sample. A frequently used measure is the interquartile range (IQR). It is calculated as the 75th percentile of the sample minus the 25th percentile of the sample. Hence the interquartile range ignores the extreme of the observations. Returning to Example 3.1, we see that the 25th percentile of the sample is equal to 0.01 and the 75th percentile of the sample is equal to 0.02, therefore IQR = 0.01. Clearly, looking only at the IQR or simple range of a sample can be very misleading. It can be argued that a respectable measure of variability should take into account for all of the observations in the sample and not for just those at the extremes. The sample variance is one measure that achieves this.

3.2.3   Sample variance and sample standard deviation

The sample variance measures how tightly individual values are clustered around the arithmetic mean of a sample. The estimator of the sample variance (S^2) is

$$S^2 = \frac{\sum_{i=1}^{N} (r_i - \bar{r})^2}{N - 1}.$$

The sample standard deviation is the square root of the above formula and is often referred to as volatility. The calculation of variance involves averaging the distance of each observation from the mean. For an individual price return r_i, the distance from the mean is measured by $r_i - \bar{r}$. For N observations in our sample we therefore have N such distances, one for each r_i. Now a property of the arithmetic mean is that

$$\sum_{i=1}^{N}(r_i - \bar{r}) = \sum_{i=1}^{N} r_i - \sum_{i=1}^{N} \bar{r} = \sum_{i=1}^{N} r_i - N\bar{r} = \sum_{i=1}^{N} r_i - N \sum_{i=1}^{N} \frac{r_i}{N}$$

$$= \sum_{i=1}^{N} r_i - \sum_{i=1}^{N} r_i = 0.$$

As a consequence the formula for variance uses the squared distances because

$$\sum_{i=1}^{N}(r_i - \bar{r})^2 \geq 0.$$

When there is little variability in the observations, they will tend to cluster close together around the arithmetic mean and therefore the sample variance and sample standard deviation will be small. The more spread out the observations are around the mean, the larger the sample variance and standard deviation. The higher the variance the more likely we will see large movements away from the mean. This is illustrated in Figure 2.5 which shows the probability distribution of two random variables with expected value equal to zero and standard deviations equal to 1 and 2 respectively. The shorter fatter curve corresponds to the random variable with the standard deviation equal to 2. If all the values in the sample take the same value they each equal the arithmetic mean. In this setting variance is equal to zero.

3.2.4   Calculating volatility for different timescales

We may wish to calculate from daily data the weekly, monthly, or yearly volatility. A simple way to achieve this is via the following rule.

Square root of time rule: $S_N = S_D \times \sqrt{\text{time}}$,

where S_N is volatility of the time-period you are considering and S_D the daily volatility/standard deviation. To illustrate this suppose the daily volatility was estimated at 2% . Since there are five trading days in a week we calculate a weekly estimate of volatility as

$$S_{\text{Weekly}} = S_D \times \sqrt{\text{time}} = 2\% \times \sqrt{5} = 4.47\%.$$

Assuming 21 days in a trading month and 252 trading days in a year, monthly and yearly estimates could be calculated as

$$S_{\text{Monthly}} = S_D \times \sqrt{\text{time}} = 2\% \times \sqrt{21} = 9.17\%,$$
$$S_{\text{Yearly}} = S_D \times \sqrt{\text{time}} = 2\% \times \sqrt{252} = 31.75\%.$$

3.3   A NORMALLY DISTRIBUTED MODEL FOR ENERGY PRICE RISK

Previously we developed the basic model of price returns as

$$R \sim N(0, 1)$$

Table 3.1 Measures of central location and variability for the log price return of Brent Crude (2-month forward), Gas Oil (spot), and the Dow Jones Swiss Electricity price index

	Brent Crude forward (%)	Gas Oil spot (%)	DJS Electricity index (%)
Median	0.07	0.00	0.00
Average	0.08	0.09	0.10
Standard deviation	1.88	2.24	17.60

Now using Table 3.1 we are in a position to specify a more precise probability model for each product

$$R_{\text{Brent}} \sim N(0.0008, 0.0188)$$
$$R_{\text{Gas Oil}} \sim N(0.0009, 0.0224)$$
$$R_{\text{Electricity}} \sim N(0.001, 0.176).$$

These models capture the differences in means and volatility between the various products and thus provide a complete characterization of uncertainty surrounding changing prices. Our only concern now is whether the normal distribution provides an adequate approximation of the distribution of uncertainty surrounding price changes in these products. Further descriptive insight into this issue can be obtained by considering the degree of asymmetry known as skew and "fat tailedness" known as kurtosis.

3.4   MEASURES OF SHAPE

Measures of central tendency and dispersion provide two important tools for summarizing a sample. They are important because they allow us to characterize the sample of price returns $\{r_1, r_2, ..., r_N\}$ by using only two metrics such as the arithmetic mean and standard deviation. In Table 3.1 we list median, arithmetic mean, and standard deviation for the daily log price return for 2-month forward Brent Crude, Gas Oil spot, and Dow Jones Swiss electricity price index over the period February 2000 to February 2004. Whilst the median and arithmetic mean for each product are close to zero they differ considerably in terms of their standard deviation, with the electricity index being more than nine times more volatile than Brent Crude. This information can be used to assist us in developing models specific to each product.

3.4.1   Coefficient of skew

The skew of a sample is a measure of the degree of asymmetry. Figure 3.1 depicts the empirical distribution for three samples on random variables. The top diagram depicts the situation where values are positively skewed. The middle diagram depicts the situation where the values are negatively skewed and the bottom diagram depicts the situation where the values are symmetrically distributed about their mean. In this example the main difference in shape between the three samples is their degree of symmetry.

The degree of skew provides information about the likelihood of extreme events. Given the sample of price returns $\{r_1, r_2, ..., r_N\}$ with sample standard deviation S and sample mean $\bar{r}$, it can be calculated as

$$\delta = \frac{\sum_{i=1}^{N}(r_i - \bar{r})^3/N}{S^3}.$$

For a symmetric distribution, large values are about as likely as small values and therefore skew is equal to zero. A negative value of skew indicates that large negative values are more likely than large positive values. A positive value of skew indicates that large positive values are more likely than large negative values. A useful rule of thumb for assessing the degree of skew is as follows:

1. If the estimated skew is greater than 1 in absolute value the distribution is highly skewed.
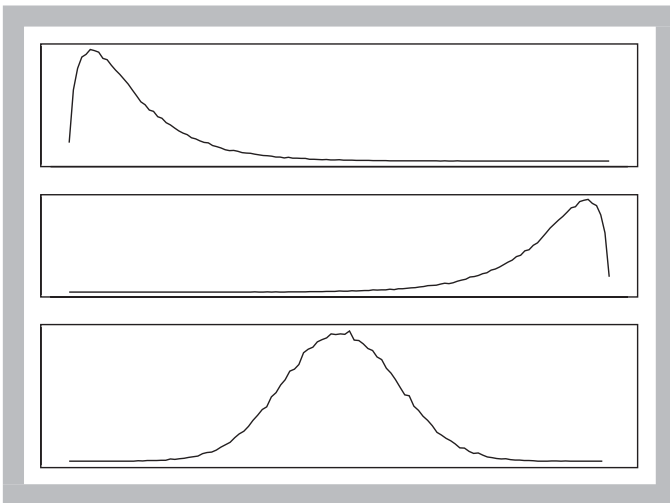
Figure 3.1 Skewed and symmetric distributions

Table 3.2 Three samples with the same sample means and standard deviations

Observation	Sample 1	Sample 2	Sample 3
1	−1.26	−1.64	1.64
2	−0.63	−0.29	0.29
3	0.00	0.60	−0.60
4	0.63	0.69	−0.69
5	1.26	0.63	−0.63
Average	**0.00**	**0.00**	**0.00**
Standard deviation	**1.00**	**1.00**	**1.00**

2. If the estimated skew lies between $1/2$ to 1 the distribution has moderately skew.

3. If the estimated skew is less than $1/2$ the distribution is fairly symmetrical.

Example 3.2   Illustrating the relevance of skew

As an illustration of the relevance of using skew to help characterize a sample of price returns, consider the three samples shown in Table 3.2. Each sample has an arithmetic mean of approximately zero and a standard deviation of approximately 1. Is this enough information to adequately characterize the sample? In general no; we also need to know something about the shape characteristics. On close careful inspection you will observe that sample 1 is symmetric around 0, sample 2 negatively skewed and sample 3 positively skewed. Using the above formula we estimate the skew for sample 1 $= 0$, skew for sample 2 $= -0.7$ and skew for sample 3 $= 0.7$. Since for a skewed distribution, the median is not equal to the mean an alternative measure of skew is

$$\delta_m = \frac{3(\text{mean} - \text{median})}{S}.$$

If (mean − median) > 0 then the sample is positively skewed. If (mean − median) < 0 then the data is negatively skewed. If (mean = median) the data is symmetric. Using this measure on Example 3.2, we find the skew for sample 1 $= 0$, skew for sample 2 $= -1.8$ and skew for sample 3 $= 1.8$. Despite the fact that the two measures of skew δ and δ_m are different estimators and hence return different numerical values they both convey the same essential information – that is, sample 1 is symmetric, sample 2 is negatively skewed and sample 3 is positively skewed.

3.4.2   Coefficient of kurtosis

Kurtosis is a measure of the size of the tails of a probability distribution. It provides information on whether the sample data are peaked or flat. Samples with high kurtosis, known as leptokurtic, tend to peak around their mean, decline rapidly, and have heavy tails. Samples with low kurtosis are known as platykurtic and tend to have a flat top near the mean. Different probability distributions have different values for kurtosis. The Normal distribution has a kurtosis of 3. Probability distributions which have the same kurtosis as the Normal distribution are known as mesokurtic. A coefficient of kurtosis can be calculated as

$$\psi = \frac{\sum_{i=1}^{N}(r_i - \bar{r})^4/N}{S^4},$$

where S is the sample standard deviation. Kurtosis is usually measured relative to the normal distribution in which case a coefficient of relative kurtosis can be calculated as:

$$\psi_{\text{Relative}} = \frac{\sum_{i=1}^{N}(r_i - \bar{r})^4/N}{S^4} - 3.$$

Example 3.3   Measuring kurtosis for three small samples

As an illustration we calculate the kurtosis of each of the three samples in Example 3.2. Since the mean for each sample is zero and the variance is equal to 1, the kurtosis is calculated as:

1. For sample 1 $\psi_1 = \frac{(-1.26)^4+(-0.63)^4+(0.00)^4+(0.63)^4+(1.26)^4}{5} = 1.07$

2. For sample 2 $\psi_2 = \frac{(-1.64)^4+(-0.29)^4+(0.60)^4+(0.69)^4+(0.63)^4}{5} = 1.55$

3. For sample 2 $\psi_2 = \frac{(1.64)^4+(0.29)^4+(-0.60)^4+(-0.69)^4+(-0.63)^4}{5} = 1.55.$

3.5   RELEVANCE OF DESCRIPTIVE STATISTICS

We can now see that descriptive statistics provide useful insight on the characteristics of a sample. Recall that, in Example 3.2, we were not able to distinguish between the three samples when we considered only their mean and standard deviation. However by collectively using the statistical measures of shape – skew and kurtosis, we were able to characterize the samples with a higher degree of accuracy. Sample 1 is symmetric with a lighter tail than samples 2 and 3. Whilst samples 2 and 3 have the same

Table 3.3 Descriptive statistics for the log price return of Brent Crude (2-month forward), Gas Oil (spot), and the Dow Jones Swiss Electricity price index

	Brent Crude forward	Gas Oil spot	DJS Electricity index
Median	0.07%	0.00%	0.00%
Average	0.08%	0.09%	0.10%
Standard deviation	1.88%	2.24%	17.60%
Skew	−0.1039	−0.3411	0.7573
Kurtosis	4.3510	4.6212	9.5213

degree of kurtosis they differ in the direction of their skew. Such information is crucially important when attempting to select an appropriate probability model for each of the respective samples.

Table 3.3 lists the descriptive statistics for the daily log price return for two month forward Brent Crude, spot Gas Oil, and Dow Jones Swiss (DJS) Electricity price index over the period February 2000 to February 2004. Both Brent Crude and Gas Oil are negatively skewed whilst DJS Electricity is positively skewed. We also see that DJS Electricity is not only considerably more volatile than the other two products, it also has more probability in the tail of its probability distribution, with a kurtosis in excess of 9. This is three times the kurtosis of a normally distributed random variable. The kurtosis of Brent Crude and Gas Oil at 4.35 and 4.62 are also slightly higher than what might be expected if they were generated from a normal distribution. Due to the large value of kurtosis the model: $R_{\text{Electricity}} \sim N(0.001, 0.176)$, appears to be an inappropriate approximation for DJS Electricity. We can check this more formally using the tools of statistical inference. Statistical inference will aid us in assessing whether the sample estimates such as the coefficient of skew or kurtosis are sufficiently different from those values we expect to see.

3.6 SUMMARY

Descriptive statistics provide summary numbers that carry information about the characteristics of a sample. The three important ways of describing a sample are:

(1) Central tendency, which indicates where the mass of observations lie.

(2) Variability which indicates how the observations are dispersed.

(3) Shape which measures the degree of symmetry and weight of probability in the tails of the empirical distribution.

Key measures of central tendency are the mean and the median. The mean is most useful when the sample is symmetric. The median is preferable when the sample is asymmetric. Variability is typically measured using variance or the standard deviation and shape through coefficients of skew and kurtosis.

3.7 FURTHER READING

Gentle introductions to descriptive statistics can be found in Crow *et al*. (1960), Rowntree (1981), Lewis (2003, 2004). Extensive discussion of alternative methods of calculating kurtosis are given in Moors (1998) and Groeneveld (1998).

Crow, E. L., Davis, F. A., and Maxfield, M. W. (1960) *Statistics Manual*, Dover Publications Inc., New York.

Groeneveld, R. A. (1998) A class of quartile measures for kurtosis. *American Statistician*, 51, 325–329.

Lewis, Nigel Da Costa (2003) *Market Risk Modeling: Applied Statistical Methods for Practitioners*. Risk Books, London.

Lewis, Nigel Da Costa (2004) *Operational Risk with Excel and VBA: Applied Statistical Methods for Risk Management*. John Wiley & Sons, Inc., New York.

Moors, J. J. A. (1998) A Quantile alternative for kurtosis, *Statistician*, 37, 25–32.

Rowntree, D. (1981) *Statistics Without Tears*, Penguin Books, Middlesex, England.

3.8 REVIEW QUESTIONS

1 Given a sample on a random which can take positive or negative values, which would you expect to be larger (mean or median) if
   (a) the observations are symmetric,
   (b) the observations are asymmetric with more large positive values than large negative values (positively skewed),
   (c) the observations are asymmetric with less large positive values than large negative values (positively skewed)?

2 Explain the difference between "sample" and "population."

3 Explain why the mean and variance alone are not necessarily adequate descriptors for a random variable.

4 For the following data:
   6.5, 7.6, 1.2, 5.6, 9.0, 8.9, 7.8, 3.4
   (a) Calculate the mean and median.
   (b) Calculate the degree of skew using all of the methods discussed in the chapter and comment on your findings.

Inferential Statistical Methods for Energy Risk Managers

Unlike descriptive statistics, inferential statistics are procedures for determining whether it is possible to make generalizations based on the data collected from a sample. Such generalizations are about an unobserved population. A population consists of all values (past and future) of the random variable of interest. In most circumstances the exact value of a population parameter such as the mean or variance will be unknown, and we will have to make some conjecture about its true value. In Chapter 3, we used sample estimators such as the mean, median, skew, and kurtosis, to provide estimates of the respective population parameters. When a sample is drawn from a population, the evidence contained within it may bolster our conjecture about population values or it may indicate that the conjecture is untenable. Hypothesis testing is a formal mechanism by which we can make and test inferential statements about the characteristics of a population. It uses the information contained in a sample to assess the validity of a conjecture about a specific population parameter.

4.1 WHAT IS A HYPOTHESIS?

A hypothesis is a simple claim or statement about a property of a population. Not only is it a statement, but it is a statement written in such a way that it can be supported or not by comparison with known facts. Researchers in all fields of science, industry, medicine, and technology will have hypotheses

about particular aspects of the populations that concern them. A hypothesis can thus be regarded as a statement about a population parameter. The following are all examples of hypotheses:

- The average time taken for the spot price of electricity to return to its long-term mean is 8 hours.

- The average price of heating oil increases during the winter months.

- The returns of the *Financial Times* 100 index are positively related to the returns in the nearest Appalachian Coal Futures contract.

Data is collected and analyzed for evidence that will either support or cast doubt on the hypothesis. The procedure by which empirical evidence from a sample of data is used to make an inferential statement about a hypothesis statement is known as a hypothesis test.

4.2   WHAT IS THE POINT OF HYPOTHESIS TESTING?

The goal of a hypothesis test is to decide whether a sample is consistent with a particular hypothesis about an unknown characteristic of the population from which the sample came. If we had full knowledge of a population, we would know what its characteristics are, and therefore whether our hypothesis was true or false. However in practice we have to make decisions based on a sample, sometimes small. Furthermore, we know that a sample estimate of a population characteristic will vary from sample to sample. To see that this is the case, we estimate the mean for six random samples generated via the statistical package R[1] each containing ten observations from a standard normal distribution:

```
>for (i in 1:5) Sample= rnorm(10,0,1)
>mean(Sample)
[1] 0.2858258
>for (i in 1:5) Sample= rnorm(10,0,1)
>mean(Sample)
[1] 0.2904667
>for (i in 1:5) Sample= rnorm(10,0,1)
>mean(Sample)
[1] 0.3330377
>for (i in 1:5) Sample= rnorm(10,0,1)
>mean(Sample)
[1] -0.1479695
>for (i in 1:5) Sample= rnorm(10,0,1)
>mean(Sample)
```

```
[1] -0.6827243
>for (i in 1:5) Sample= rnorm(10,0,1)
>mean(Sample)
[1] -0.5393777
```

In the first sample the mean is 0.2858258, in the second it is 0.2904667, and in the tenth sample it is −0.5393777. Even though the samples are from exactly the same probability distribution we see that only by chance will we obtain the exact sample estimate for differing samples. Indeed, only by chance will a sample estimate equal exactly the population characteristic. The example illustrates the important point that a sample only provides approximate knowledge of the population, which may or may not be close enough to the true characteristics to lead us to make the correct inference. Intuitively, the larger the sample the more confidence we may have in the sample estimates. We can see that our intuition is indeed correct by re-running the above six samples but this time with a much larger sample size of 10,000 observations:

```
>for (i in 1:5) Sample= rnorm(10000,0,1)
>mean(Sample)
[1] -0.009000733
>for (i in 1:5) Sample= rnorm(10000,0,1)
>mean(Sample)
[1] -0.006842981
>for (i in 1:5) Sample= rnorm(10000,0,1)
>mean(Sample)
[1] -0.01050050
>for (i in 1:5) Sample= rnorm(10000,0,1)
>mean(Sample)
[1] -0.006001264
>for (i in 1:5) Sample= rnorm(10000,0,1)
>mean(Sample)
[1] -0.0006383873
>for (i in 1:5) Sample= rnorm(10000,0,1)
>mean(Sample)
[1] 0.001146425
```

In this case, the sample estimates of the mean are much closer to the population value of 0. The general rule of thumb is the larger the sample size the better.

4.3   REFUTING CHANCE AND CONTROLLING ERRORS

How can we have confidence that our sample estimate reflects the actual value of the population characteristic rather than some rogue value,

especially when we have a small sample? The answer will need to provide us with evidence on whether our estimate is in some sense significantly different from a hypothesized value. The way this problem is approached through statistical hypothesis testing is first to form two mutually exclusive hypothesis statements known as the null hypothesis (H_0) and the alternative hypothesis (H_A); and second to calculate how often we would get a sample estimate as large or larger than the observed sample estimate if the population parameter of interest really was equal to the null hypothesis and therefore the sample estimate was due to chance. If a value as large or larger than the estimate occurs by chance frequently, then chance is a feasible explanation of the observed value. However, if such a value would occur by chance very rarely, then chance is deemed not to be a feasible explanation. In this approach the assumption is that the null hypothesis is correct and our goal is to reject the null hypothesis in favor of the alternative hypothesis. The null hypothesis is always chosen to be the hypothesis in which there is no change.

As we are dealing with sample estimates there is always a possibility that our hypothesis test leads us to incorrectly reject the null hypothesis when it is in fact true. An error of this kind is known as a type I error. The probability of committing a *type I error* is known as the significance level of a hypothesis test and denoted by α. The level of significance is set by the researcher and reflects the acceptable error threshold for rejecting the null hypothesis when the null hypothesis is in actual fact true. In general it is chosen to be small typically with values of 0.01, 0.05, or 0.1 often reported as the 1%, 5%, or 10% significance level. The use of a 10% significance level, implies that we make a type I error 10% of the time. Acceptance of the null hypothesis when it is in fact false is known as a *type II error* denoted by β. Of course we would like to make the probability of making a type II error as small as possible. The *power* of a statistical hypothesis test measures the test's ability to reject the null hypothesis when it is actually false. It is calculated as $1 - \beta$. The maximum power a test can have is 1, the minimum is 0.

4.4 A STEP BY STEP GUIDE TO CONDUCTING A HYPOTHESIS TEST

The easiest way to become familiar with hypothesis testing is to work through an example. We shall illustrate the steps involved using the 2-month Brent Crude forward price return of Table 3.3. Recall the kurtosis was 4.3510. This value is somewhat larger than we might expect if the random variable was from a normal distribution. The relative kurtosis is measured with the variable RelKurt where

$$\text{RelKurt} = 4.3510 - 3 = 1.3510.$$

4.4.1 Step 1: Determine the null hypothesis

If the 2-month Brent Crude contract were normally distributed we might expect RelKurt to be equal to zero. We therefore investigate the null hypothesis that the population relative kurtosis is equal to zero against the alternative that it is greater than zero. You may be wondering why we use "kurtosis equal to zero" as the null hypothesis rather than "kurtosis greater than zero." This is because the null hypothesis is always the hypothesis that includes any of the equalities $=$, $\geq$, or $\leq$. Differentiating between the null and alternative hypothesis can be easily achieved if you remember this point.

4.4.2 Step 2: Determine whether a one-sided or two-sided test is required

Denote the sample estimate of relative kurtosis RelKurt by $\hat{\kappa}$ and the population relative kurtosis by κ. Let c be the value we are interested in testing. There are three possibilities for specifying our hypothesis test:

(a) Test H_0: $\kappa = c$ against H_A: $\kappa > c$

(b) Test H_0: $\kappa = c$ against H_A: $\kappa < c$

(c) Test H_0: $\kappa = c$ against H_A: $\kappa \neq c$

Tests (a) and (b) are known as *one-sided* tests because we are only interested in values greater than or less than the null hypothesis. Test (c) is a two-sided test because we are interested in values greater or less than the null hypothesis. In this case we are interested in values greater than zero, so we set $c = 0$. The hypothesis is one sided:

Test H_0: $\kappa = 0$ against H_A: $\kappa > 0$

4.4.3 Step 3: Set the level of significance

The probability of committing a type I error is known as the level of significance. This is the criterion for rejection or acceptance of the null hypothesis. We know the probability of a large difference between the sample relative kurtosis and population relative kurtosis will diminish as the sample size increases. For very large samples sizable differences between the population kurtosis and sample kurtosis are possible but unlikely. In setting the level of significance we are explicitly setting the probability of erroneously rejecting the null hypothesis when it is true. We choose $\alpha = 5\%$, because this is the level commonly chosen in science, academia, and industry.

4.4.4   Step 4: Calculate the test statistic

A test statistic, which we denote by $\hat{T}$, is a function of the sample data. The value estimated from the data is used to decide whether or not to reject the null hypothesis. There are numerous test statistics. For relative kurtosis, the test statistic is:

$$\hat{T} = \frac{\hat{\kappa}}{\sqrt{24/N}} = \frac{1.3510}{\sqrt{24/522}} = 6.30$$

where N is the number of observations in the sample, in this case 522.

A test statistic $\hat{T}$ is a random variable because the value it takes will vary from sample to sample. Since it is a random variable it must also have an associated probability distribution. What is the probability distribution of $\hat{T}$? In this case $\hat{T}$ is a standard normally distributed random variable. Therefore, the mean of the distribution of this test statistic is zero and the standard deviation is one. The probability distribution of a test statistic is usually referred to as the sampling distribution. The most common sampling distribution for a test statistic is the standard normal distribution. Other sampling distributions include the Student t-distribution, F-distribution and Chi-squared distribution. Probability tables for these distributions are given in Appendix 1.

4.4.5   Step 5: Obtain critical value of the test statistic and/or the *p*-value

Given a test statistic and knowledge of its sampling distribution we can calculate the critical value of the test statistic denoted T_α. This is that value beyond which the null hypothesis becomes untenable. The value of T_α is obtained from the percentile function of the test statistic $F_{\hat{T}}^{-1}(1 - \alpha)$. As in this case we have set $\alpha = 5\%$ and we know the test statistic is a standard normal random variable we find using Table A.1:

$$F_{\hat{T}}^{-1}(1 - \alpha) = 1.64.$$

4.4.6   Step 6 (a): Compare value of test statistic to critical value

Finally, we compare the value of the test statistic to the critical value and use this comparison to decide whether or not to reject the null hypothesis. As we are carrying out a one-sided test we reject the null hypothesis in favor of the alternative hypothesis if $\hat{T} > T_\alpha$. Since $\hat{T}$ (which is equal to 6.30) is

greater than $T_{0.05}$ (which is equal to 1.64) we reject the null hypothesis of no excess kurtosis.

If the null hypothesis is true then values of the test statistic $\hat{T}$ near zero are much more likely than values far away from zero. The null hypothesis is rejected if and only if the evidence against it is strong in the sense that the estimate of $\hat{T}$ could not easily have been produced by chance. Since, in the above illustration, the probability of observing a value of 6.30 from a standard normal distribution is close to zero we have little choice but to reject the null hypothesis. When the null hypothesis is rejected in favor of the alternative hypothesis the result is referred to as statistically significant.

The p-value of a test statistic is:

$$p\text{-value} = 1 - F(\hat{T}),$$

where $F(.)$ is the cumulative distribution function of the sampling distribution of the test statistic. This provides an equivalent approach to testing the null hypothesis – reject the null hypothesis if the p-value is less than the significance level. In other words if $\hat{T} > T_\alpha$ then the p-value of $\hat{T}$ is less than α.

4.4.7   Step 6 (b): Two-sided test

If we specified our hypothesis as

Test $H_0: \kappa = 0$ against $H_A: \kappa \neq 0$,

this would be a two-sided test. In this case we reject the null hypothesis if $|\hat{T}| > T_\alpha$. The only adjustment to the calculations is that we calculate the critical value using $\alpha/2$. For example, if we set the level of significance at $\alpha = 5\%$ and our test statistic is standard normally distributed, the critical value of the test statistic in a two-sided hypothesis test is $T_{0.025} = 1.96$ (see Table A.1).

4.4.8   A test statistic for the mean

We have already seen that the sample mean is one useful way of summarizing information contained in a sample. Often we will be interested in knowing whether the sample estimate is significantly different from some predetermined level. Given a small sample of thirty or less observations on a random variable we may be asked to comment on whether the average of the observations is significantly different from some value μ_0. If we denote the sample mean by $\bar{x}$ and sample standard deviation of S the test statistic

is given by:

$$\hat{T} = \frac{(\bar{x} - \mu_0)}{S/\sqrt{N}}.$$

Example 4.1   Testing a hypothesis about a sample mean

Suppose a risk manager has collected the following observations on the year-end value of the oil gas spot price and is interested in testing whether the price mean is greater than 230

Year	1998	1999	2000	2001	2002	2003	2004
Reported	$270	$273	$258	$204	$228	$282	$254

The sample mean is equal to $252.7143, which appears to be somewhat larger than $230. The sample standard deviation is equal to $27.6328.

As we only have a handful of observations we can use the previously mentioned test statistic. We are interested in values above $230; the null and alternative hypothesis are

Test H_0: population mean $\leq$ $230

against

H_A: population mean > $230.

This is a one-sided test. We set $\alpha = 5\%$. The value of the test statistic $\hat{T}$ is

$$\hat{T} = \frac{(252.7143 - 230)}{(27.63280/\sqrt{7})} = 2.1748.$$

The test statistic has a t-distribution with 7 degrees of freedom and is equal to 1.895 (see Table A.3).

Hence, the critical value of the test statistic is $T_{0.05} = 1.895$. Since the test statistic $\hat{T}$ (equal to 2.1748) is greater than the critical value $T_{0.05}$ we reject the null hypothesis. The p-value (probability that the test statistic is greater than or equal to $\hat{T}$) is given by $F_{\hat{T}}^{-1}(1 - \alpha)$ (where $F_{\hat{T}}^{-1}(\ )$ is now the percentile function of the t-distribution with 7 degrees of freedom) and equal to 0.03629. If the p-value is less than the level of significance (0.05) we reject the null hypothesis. This is indeed the case in this example and we therefore conclude that there is sufficient evidence to support the assertion that the average price of Gas oil is statistically significantly higher than $230.

4.5   CONFIDENCE INTERVALS

A confidence interval is an interval constructed from a sample, which includes the parameter being estimated with a specified probability known as the level of confidence. The level of confidence is simply one minus the level of significance. Therefore, a significance level of 5% implies a 95% confidence level. A confidence interval provides information on the range of plausible values in which the population parameter is likely to lie. Most statistical packages include an option for calculating confidence intervals. For example, the statistical package R provides the function:

```
norm.ci(t0=parameter, conf=confidence level,
var=sample variance)
```

An alternative, which can be easily entered into a spreadsheet package is to use the formula

$$\text{Confidence interval} = \text{sample mean} \pm F_{\hat{T}}^{-1}(1 - \alpha)$$

$$\times \text{sample standard deviation,}$$

where $F_{\hat{T}}^{-1}(1 - \alpha)$ is the percentile function of the standard normal distribution.

For example, if we use the above formula to construct a 95% confidence around the mean (0.08%) of the daily return of Brent Crude we obtain a confidence interval which covers the range from -3.60% to 3.76%. What does this tell us? First, it specifies a plausible range of values within which the unknown population mean may lie – the actual (population) value is likely to lie somewhere close to this value say between -3.60% to $+3.76\%$. Second, since a confidence interval is a function of a sample it is itself a random variable and will therefore vary from sample to sample. Third, if we were to calculate the sample mean from many different samples, and the confidence interval is calculated for each sample then, 95% (since we used 95% confidence level in our calculation) of such intervals would cover the true population mean. The width of the confidence interval measures how uncertain we are about the unknown population parameter. A narrow interval indicates less uncertainty about the value of the population parameter than a wide interval.

4.6   SUMMARY

Statistical inference offers a formal framework by which we can assess the significance of a hypothesis or parameter from a sample. Its importance

in risk management lies in its ability to provide confirmatory or otherwise proof about quantitative issues of importance. It can thus be utilized by the risk manager to provide additional empirical evidence in support or against a particular argument or idea.

4.7   FURTHER READING

A through introduction to statistical inference can be found in Crow *et al.* (1960). An alternative aimed specifically at risk managers is Lewis (2003). For a more gentle overview, see Rowntree (1981). Lewis (2004) offers a number of applied examples relevant for risk management.

Crow, E. L., Davis, F. A. and Maxfield, M. W. (1960) *Statistics Manual*, Dover Publications Inc., New York.
Lewis, Nigel Da Costa (2003) *Market Risk Modeling: Applied Statistical Methods for Practitioners*, Risk Books, London.
Lewis, Nigel Da Costa (2004) *Operational Risk with Excel and VBA: Applied Statistical Methods for Risk Management*, John Wiley & Sons, Inc., New York.
Rowntree, D. (1981) *Statistics Without Tears*, Penguin Books, Middlesex, England.

4.8   REVIEW QUESTIONS

1 If a hypothesis test fails to reject the null hypothesis we accept it – Discuss.

2 What is the difference between a type I error and a type II error?

3 Which do you feel is more useful – a hypothesis test or a confidence interval and why?

Applied Modeling: Techniques and Applications

I was once taught by a truly great mathematics professor, who in a private moment, scolded me for using plain(ish) English to describe my new solution to an old mathematical/statistical problem. He suggested I wrap my discovery up in an abstract and obscure fashion, so as to keep the technique from falling into the hands of those who might find a use for it. In a slow and disapproving voice he lamented:

> God forbid anything produced by one of MY students should prove to be of any practical value! Solving applied problems and providing useful insight into the actual workings of the world should be left to ... to Social Scientists or failing that applied Mathematicians, Econometricians, Engineers and Physicists. It should never ever be attempted by one of my students.

I was politely but firmly informed that I should spend more of my days extolling the beauty of ultra complex equations whilst bombarding my fellow students, friends, family, and anybody else who happened by, with barely intelligible mathematical mutterings, abstract theories and thought-provoking conjectures. If I was to become "one of the lads" I would have to put more effort into proliferating the erudite haze that shrouds mathematics. For a while I played along. I even purchased a secondhand tweed jacket with matching corduroy trousers. However, the thought of my metamorphosis from poorly paid lecturer in moth-eaten clothes into a wide-eyed wild haired "mad" professor proved not very appealing and I became "rocket scientist" for a large financial firm.

The attitude of "*don't touch me with the dirty facts of reality – they get in the way of my beautiful equation*" still persists in much of the academic quantitative community; It can even be found by some (generally academics) inside the discipline of risk management. However, The plain fact of the matter is that financial risk management is a practical discipline; Risk managers have to deal with real people, inconvenient facts, messy data, make unsatisfactory assumptions, and often simply guess. In other words they, unlike my former professor, have to get their hands dirty. We too will have to get our hands dirty. For this section of the book outlines the tools necessary for thinking about and exploring energy price risk from a statistical perspective. The techniques are not always completely satisfactory from a theoretical perspective, but those discussed in this section are intuitive and work reasonably well in practice. We begin, in Chapter 5, by describing an approach for fitting probability distributions to energy price returns. Chapter 6 discusses nonparametric density estimation as an alternative to parametric specifications. Chapter 7 introduces correlation analysis, followed in Chapters 8–11 with topics surrounding the theory and use of regression models. Chapter 12 discusses various standard approaches to modeling volatility. Finally Chapter 13 introduces and discusses a number of stochastic differential equations, the building blocks used in pricing energy derivatives.

Modeling and Fitting Price Distributions

Throughout the energy sector, risk managers, and analysts face the challenge of uncovering the price and return distributions of various products. Knowledge about the underlying probability distributions generating returns is used both in pricing models and risk management. The selected probability distribution(s) can have a significant impact on the calculated Value at Risk measure of a company's exposure from trading floor transactions and in the use of derivative pricing tools. It is imperative that risk management metrics such as Value at Risk are calculated using a statistical distribution tailored to the specific characteristics of the energy product of interest. Fitting probability distributions by carefully analyzing energy price returns is an important, although often neglected, activity. This may be partly because the number and variety of distributions to choose from is very large. For a specific product such as the forward price of Brent Crude, or price return of an Electricity index, which of the dozens of distributions should we use? This chapter outlines the process by which the practicing risk manager can begin to answer this question. It starts by assessing the validity of a simple model based on the normal distribution. When normality fails we can adjust the percentiles of the normal probability distribution. If this does not appear to help we might select an alternative probability distribution or else consider a mixture of normal distributions.

5.1   DEVELOPING A SIMPLE MODEL FOR ENERGY RETURNS

Ideally, when it comes to statistical modeling, we want a things to be as simple as possible. Simplicity means, provided we are satisfied with

the explanatory power, we should make our models as parsimonious as possible. Simplicity also aids us in our duty to explain assumptions, models, and implications to co-workers and senior management. The easiest assumption to make about price returns is to assume they are normally distributed

$$R \sim N(\mu, \sigma).$$

If this is so, our analysis can be greatly simplified, because estimates of the unknown parameters μ and σ can be obtained as

$$\hat{\mu} = \frac{r_1 + r_2 + \cdots + r_N}{N},$$

and

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{N}(r_i - \hat{\mu})^2}{N-1}}.$$

These two equations should feel familiar, they are the same as those used to calculate the sample mean and sample standard deviation respectively. Provided this model satisfactorily describes the price return of the energy contract of interest we should not shy away from using it. To aid our discussion we rewrite the above model slightly modified as

$$r_t = \mu + \sigma \times \varepsilon_t, \quad \text{where } \varepsilon_t \sim N(0, 1).$$

The model tells us that on average we can expect a return of μ with standard deviation equal to σ. The observed return r_t will differ from this value primarily due to the random variable ε_t which has a standard normal distribution. In the case of daily returns of equity markets such as those listed on the London Stock Exchange or the New York Stock Exchange it is often reasonable to assume $\mu = 0$ with σ somewhere in the range of 0.10 to 0.4. Indeed over the period April 1994 to end of March 2004, σ for various world equity indices varied widely at 0.145, 0.152, 0.173, 0.215, and 0.286 for the FTSE 100, S&P 500, Nikkei 300, French CAC 40, and Hong Kong Hang Seng index respectively.

In Chapter 3 the daily price returns of the 2 month forward Brent Crude, Gas Oil spot and Dow Jones Swiss electricity price index were modeled as

$$R_{\text{Brent}} \sim N(0.0008, 0.0188)$$

$$R_{\text{Gas Oil}} \sim N(0.0009, 0.0224)$$

$$E_{\text{Electricity}} \sim N(0.001, 0.176)$$

Therefore, for Brent Crude we can write

$$r_t = 0.0008 + 0.0188 \times \varepsilon_t,$$

whilst for Gas Oil over the same period we might specify

$$r_t = 0.0009 + 0.0224 \times \varepsilon_t,$$

and for the Dow Jones Swiss Electricity returns

$$r_t = 0.001 + 0.176 \times \varepsilon_t.$$

In all three cases we assume $\varepsilon_t \sim N(0, 1)$.

Having specified a model for price returns, we will need to assess its adequacy. How should we do this? Well, it requires first that we make explicit the assumptions underlying our postulated model. The most obvious assumption is about the distribution of ε_t and we investigate this further in the remaining pages of the chapter. The other assumptions, which may not be so obvious concern the nature of μ and σ, we shall explore these further later in this chapter and through much of the rest of this text.

5.2   USING DESCRIPTIVE STATISTICS TO ASSESS THE MODEL

We have discussed descriptive statistics in detail in Chapter 3. We already know that a normally distributed random variable is symmetric, with the mean equal to the median, a coefficient of skew equal to zero and kurtosis equal to 3. Sample values much different from these, indicate the possibility that the sample is not from a Normal distribution. Table 5.1 shows some descriptive statistics for the standardized daily log price return of the

Table 5.1 Descriptive statistics for the standardized log price return of Brent Crude (2-month forward), Gas Oil (spot), and the Dow Jones Swiss Electricity price index

	Brent Crude	Gas Oil	DJS Electricity
Median	−0.005	−0.039	−0.006
Average	0.000	0.000	0.000
Standard deviation	1.000	1.000	1.000
Skew	−0.1039	−0.3411	0.7573
Kurtosis	4.3510	4.6212	9.5213

2-month forward Brent Crude, Gas Oil (spot), and Dow Jones Swiss Electricity price index over the period February 2000 to February 2004. Notice that for all three products there is only a small degree of skew, which is also reflected in the smallish difference between the sample median and the sample mean. Since the degree of skew is rather mild,[1] we might not be too phased in using a symmetric probability distribution such as the Normal distribution. However, the high degree of kurtosis is of some concern, particularly in the Electricity index, which at 9.52, is considerably in excess of what we might expect if ε_t were from a Normal distribution.

It is always useful to plot data. Figure 5.1 shows the histogram of the standardized returns for each of the products. All three of the histograms confirm our previous finding of approximate symmetry in the daily returns. They also indicate fatter tails than we might expect for a normally distributed random variable. The tails of the empirical distributions can be seen more clearly in the dot plots of Figures 5.2, 5.3, and 5.4.

A useful complement to a histogram and dot plot is a QQ plot – a plot of the quartiles (or %) of points below a given value of the sample data

Figure 5.1 Histogram of the daily standardized log price return of Brent Crude (2-month forward), Gas Oil (spot), and the Dow Jones Swiss Electricity price index
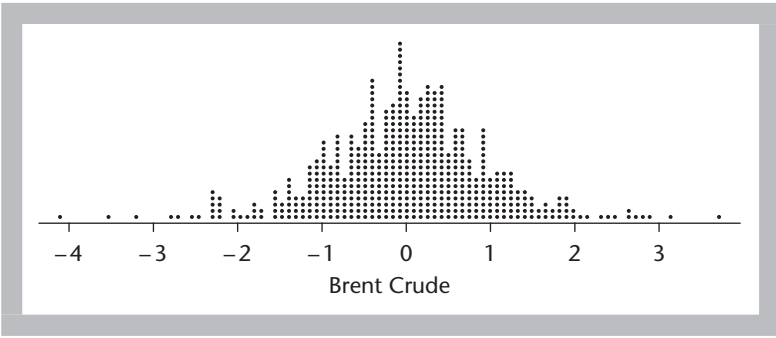
Figure 5.2 Dot plot of the daily standardized log price return of Brent Crude (2-month forward)
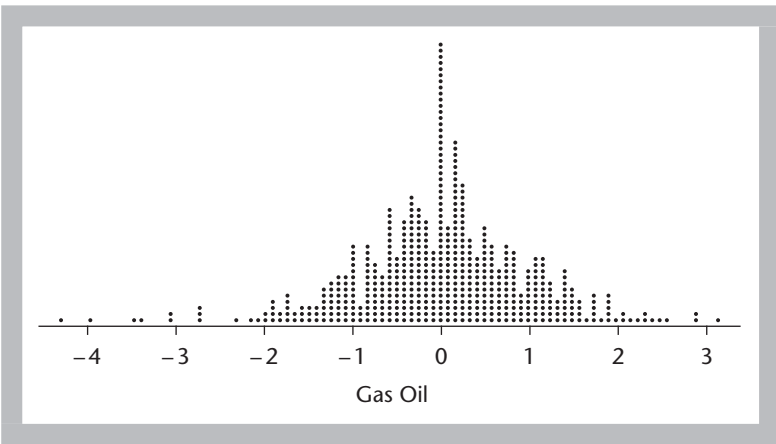
Figure 5.3 Dot plot of the daily standardized log price return of Gas Oil (spot)

set against the quartiles of the postulated probability distribution – in this case the normal distribution. A straight line (known as the reference line) is also plotted. If the sample comes from the postulated probability distribution, the plotted points will fall along this reference line. Departures from the reference line indicate departures from the specified distribution. Figure 5.5 shows a QQ plot for data which come from a standardized normal distribution. Notice how most of the points lie along the reference line. In contrast Figure 5.6 shows the situation where the data points come from an alternate (not normal) distribution. Notice how the points deviate away from the reference line, especially in the tails (high and low percentiles). Figures 5.7, 5.8, and 5.9 show the QQ plots for each of the above products. Neither of the graphs is particularly favorable to our assumption of normality. All three products exhibit significant deviations from the reference
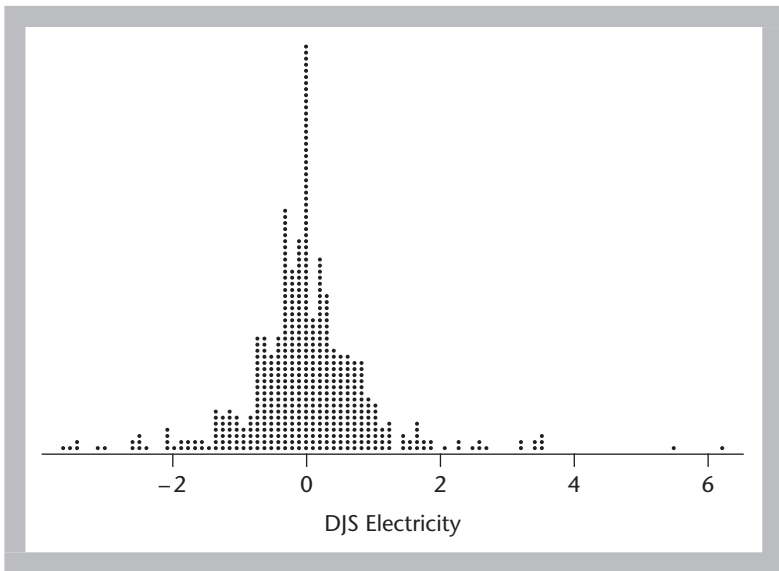
Figure 5.4 Dot plot of the daily standardized log price return of the Dow Jones Swiss Electricity price index
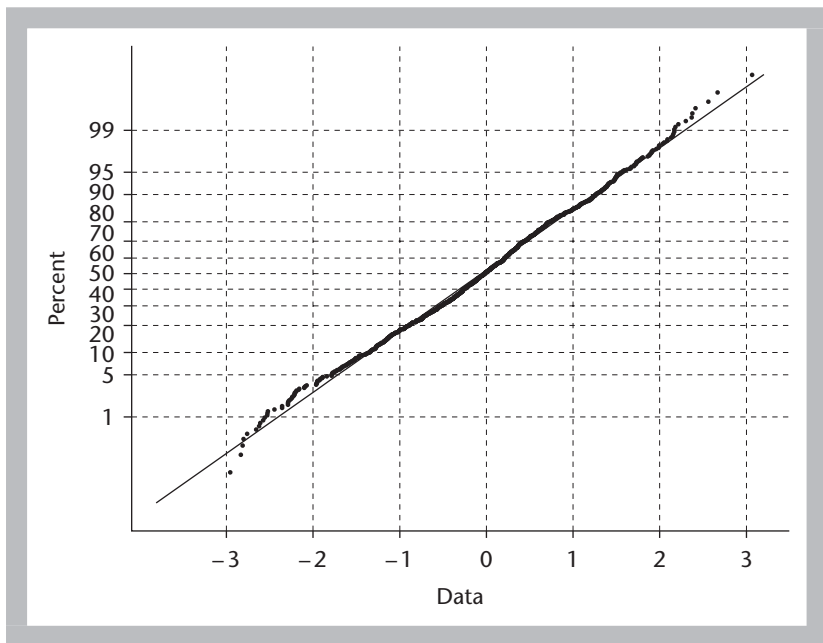
Figure 5.5 Normal distribution QQ plot for a normally distributed random sample
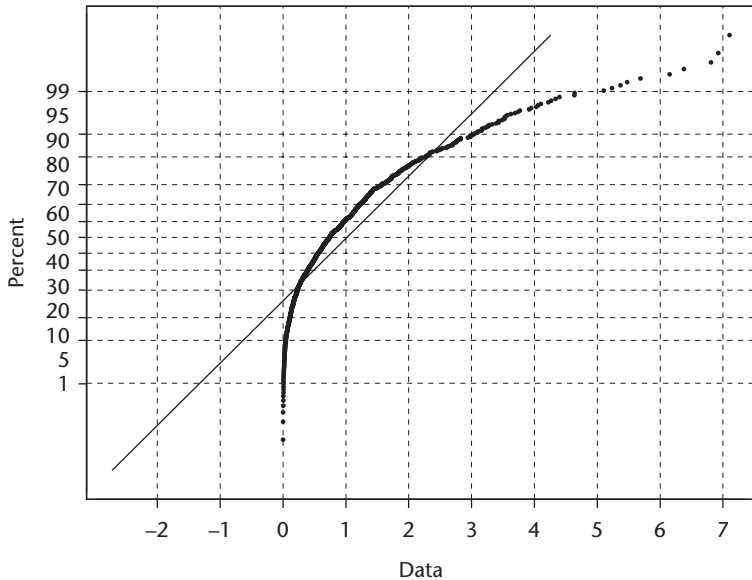
Figure 5.6 Normal distribution QQ plot for a non-normally distributed random sample
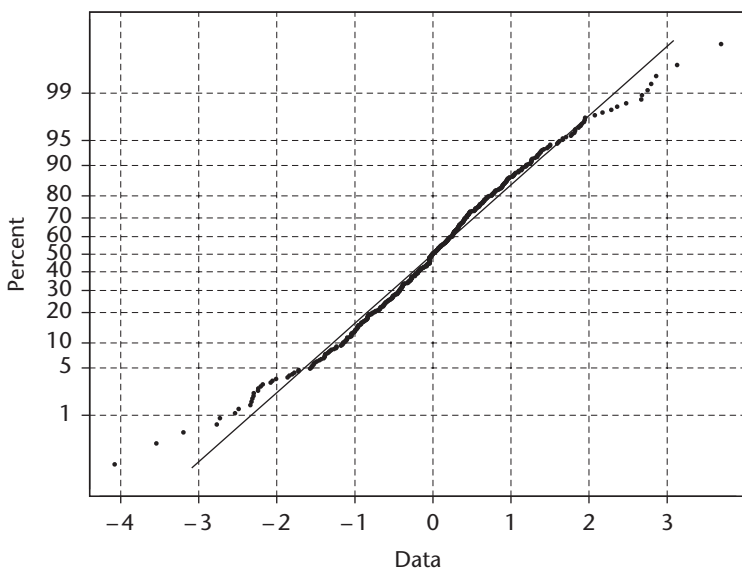
Figure 5.7 Normal distribution QQ plot of daily standardized log price return of Brent Crude (2-month forward)
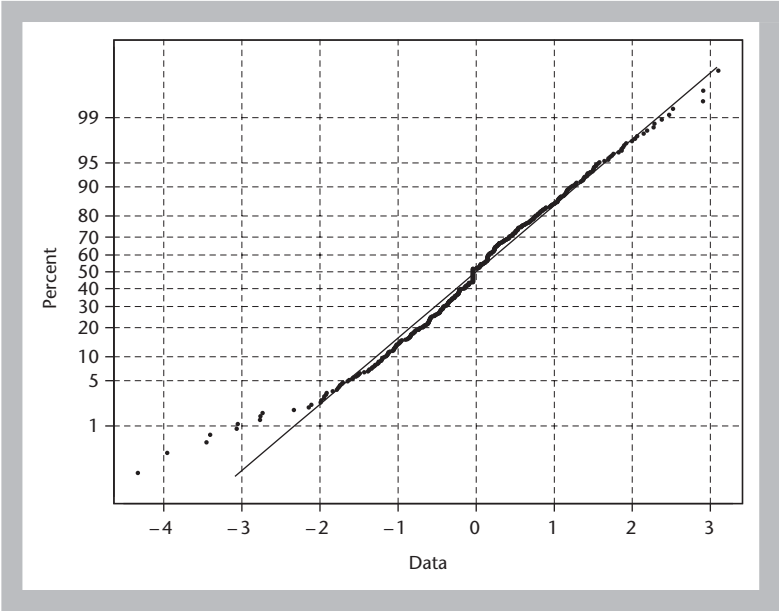
Figure 5.8 Normal distribution QQ plot of daily standardized log price return of Gas Oil (spot)
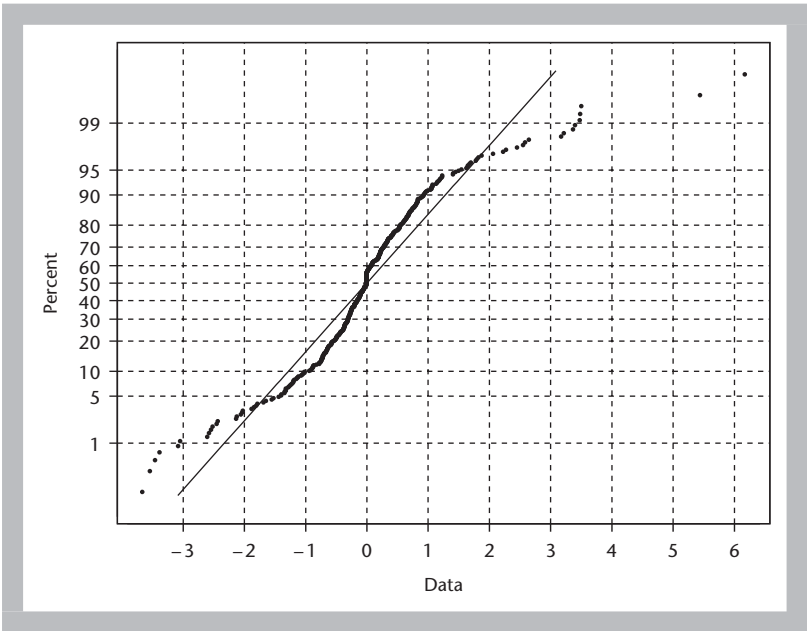
Figure 5.9 Normal distribution QQ plot of daily standardized log price return of the Dow Jones Swiss Electricity price index

lines especially in their lower tails; indeed Electricity deviates considerably in almost every respect including both upper and lower tails. The evidence against the assumption of Normality of ε_t is mounting. Nevertheless it is still worthwhile assessing the situation using a formal statistical test.

5.3   USING INFERENTIAL STATISTICS TO AID MODEL CONSTRUCTION

One approach to testing the assumption of normality would be to specify individual test statistics to assess whether the sample skew or sample kurtosis are statistically significantly different from the normal distribution. Indeed in Chapter 3 we gave a test statistic for kurtosis. An alternative is to use test statistics aimed specifically at assessing the overall degree of normality of a sample. It turns out that there are numerous test statistics for investigating this hypothesis. Two of the most widely used (and available in almost every major statistical package) are the Kolmogorov–Smirnov Test and the Anderson–Darling Goodness of Fit test.

5.3.1   The Kolmogorov–Smirnov test

The Kolmogorov–Smirnov test statistic is calculated as the largest absolute deviation between the sample cumulative distribution function (denoted by $S_N(r)$) and postulated cumulative probability distribution function (denoted by $F(r)$ and in this case the cumulative normal distribution) over the entire range of the price return random variable R:

$$\hat{T} = \max \quad |S_N(r) - F(r)| \quad \text{over all } r.$$

It can be interpreted as the maximal difference between the empirical and theoretical cumulative probability distributions. A very small difference indicates that the empirical and theoretical distributions are similar. The test relies on the observation that the sample cumulative density function is approximately normally distributed for large samples. Hence the test is distribution free in the sense that the critical values do not depend on the specific probability distribution being tested. For a 10% significance level the critical value for the test statistic is approximately $1.224/\sqrt{N}$. For 5% significance level it is approximately $1.358/\sqrt{N}$, and for a 1% significance level it is approximately $1.628/\sqrt{N}$.

5.3.2   The Anderson–Darling test

The Anderson–Darling test statistic is given by:

$$\hat{T} = -N - \frac{1}{N} \sum_{i=1}^{N} (2 \times i - 1)[\ln F(\check{r}_i) + \ln(1 - F(\check{r}_{N+1-i}))],$$

where $\check{r}_i$ are the observations ordered by size.

This test statistic is a modification of the Kolmogorov–Smirnov test and is more sensitive to deviations in the tails of the postulated probability distribution. It achieves this added sensitivity by making use of the specific postulated distribution in calculating critical values. Unfortunately this extra sensitivity comes at the minor cost of having to calculate critical values for each postulated distribution. If we postulate a normal distribution for our data then for a 10% significance level the critical value is approximately 0.631, for 5% significance level it is 0.752, and for a 1% significance level it is 1.035. However, if we postulate another distribution, say a Weibull distribution then the critical value for a 10% significance level is 0.637, for 5% significance level it is 0.757, and for a 1% significance level it is 1.038.

Table 5.2 give the Anderson–Darling test for each of the energy products in question. The p-values are all less than 0.001, and thus we conclude that each of the products exhibits significant departure from normality. We obtain similar results when using the Kolmogorov–Smirnov test. Surprised? You should not be as we were already aware of the high kurtosis of each of these products relative to that of the normal distribution. Indeed the source of the non-normality of the products appears to lie in their high relative kurtosis.

5.3.3   Jarque–Bera test

Another frequently used test of normality known as the Jarque–Bera (JB) test uses both the coefficient of skew and coefficient of kurtosis to jointly assess

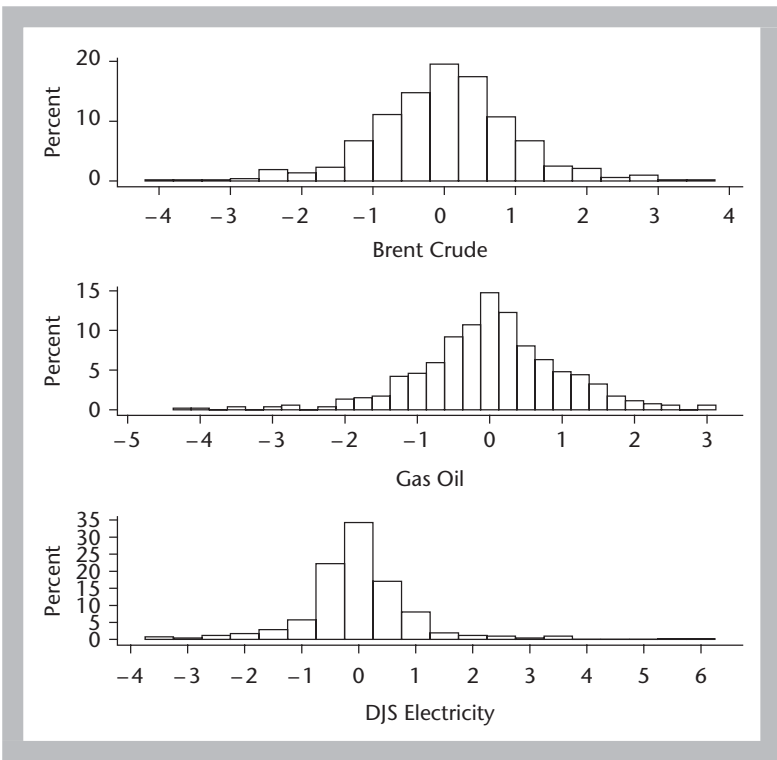**Table 5.2** Anderson–Darling test statistic for the standardized log price return of Brent Crude (2-month forward), Gas Oil (spot), and the Dow Jones Swiss Electricity price index

	Brent Crude	Gas Oil	DJS Electricity
Test statistic	1.932	2.471	14.416
p-value	Less than 0.001	Less than 0.001	Less than 0.001

Table 5.3 Jarque–Bera test statistic for the standardized log price return of Brent Crude (2-month forward), Gas Oil (spot), and the Dow Jones Swiss Electricity price index. Critical value at 5% is 5.99

Brent Crude	Gas Oil	DJS Electricity
40.64	67.29	974.86

the assumption of normality. The JB test statistic is given by:

$$\hat{T} = N\left[\frac{\delta^2}{6} + \frac{(\kappa - 3)^2}{24}\right],$$

where δ is the sample coefficinet of skew and κ is the sample coefficient of kurtosis. This test statistic has a chi-squared distribution on 2 degrees of freedom.[2] It is an inherently pleasing test statistic because it is a joint test of the null hypothesis (of normality) that sample skewness equals 0 and sample kurtosis equals 3. For a 5% significane level the critical value is 5.99 (see Table A.2). Table 5.3 lists the value of the JB test statistic for each of the products. Notice that for all of the products considered, the value of the test statistic is greater than the critical value, and hence we continue to reject the null hypothesis.

5.4 WHAT TO DO WHEN NORMALITY FAILS?

Given the available evidence we can reject the normal distribution as a probabilistic mechanism for generating price returns (at least for those specific products considered so far), right? Not exactly ... but provided we are happy to stick with the simple model $r_t = \mu + \sigma \times \varepsilon_t$, we would feel uncomfortable with the assumption that ε_t is normally distributed. The question is what to do next. We shall consider three solutions: in the first, we stick with the normal distribution but modify its percentile function; the second involves selecting an alternative probability distribution; and in the third approach we use a mixture of normal distributions.

5.4.1 Adapting the simple model via a Cornish–Fisher approximation

In the circumstances where the departure from normality appears to be relatively mild we can adjust the percentiles of the normal probability

distribution to take into account any skew and or kurtosis. Such an adjustment is known as the Cornish–Fisher approximation. The Cornish–Fisher approximation calculates the inverse probability function $F^{-1}(1 - \alpha)$ as

$$
\begin{aligned}
F^{-1}(1 - \alpha) = F^{-1}_{\text{Normal}}(1 - \alpha) + \frac{1}{6} \left(\left[F^{-1}_{\text{Normal}}(1 - \alpha) \right]^2 - 1 \right) \\
\times \delta + \frac{1}{24} \left(\left[F^{-1}_{\text{Normal}}(1 - \alpha) \right]^3 - 3 \times F^{-1}_{\text{Normal}}(1 - \alpha) \right) \times \kappa \\
- \frac{1}{36} \left(2 \times \left[F^{-1}_{\text{Normal}}(1 - \alpha) \right]^3 - 5 \times F^{-1}_{\text{Normal}}(1 - \alpha) \right) \times \delta^2,
\end{aligned}
$$

where $F^{-1}_{\text{Normal}}(1 - \alpha)$ is the percentile function of the standard normal distribution, δ is the sample skew and κ is the sample relative kurtosis. In terms of our simple model, a Cornish–Fisher adjustment implies that

$$
\tilde{\varepsilon}_t \sim N_{\text{CFA}}(0, 1),
$$

where $N_{\text{CFA}}(0, 1)$ is the Cornish–Fisher adjustment to the standard normal distribution.

If the distribution is symmetric with zero relative kurtosis then $N_{\text{CFA}}(0, 1)$ is identical to $N(0, 1)$ and hence $\tilde{\varepsilon}_t = \varepsilon_t$.

Figures 5.10, 5.11, and 5.12 plot the Cornish–Fisher adjusted (continuous line) alongside the histogram of standardized returns for Brent Crude, Gas Oil, and DJS Electricity respectively. Visual inspection of these figures reveals that the Cornish–Fisher adjusted probability distribution fits all three products much better than the standard normal distribution. Since the approximation appears reasonable we can adjust our simple model. For Brent Crude we have

$$
r_t = 0.0008 + 0.0188 \times \tilde{\varepsilon}_t,
$$

whilst for Gas Oil we would specify

$$
r_t = 0.0009 + 0.0244 \times \tilde{\varepsilon}_t,
$$

and for the DJS Electricity

$$
r_t = 0.001 + 0.176 \times \tilde{\varepsilon}_t.
$$

In all three cases we assume $\tilde{\varepsilon}_t \sim N_{\text{CFA}}(0, 1)$.
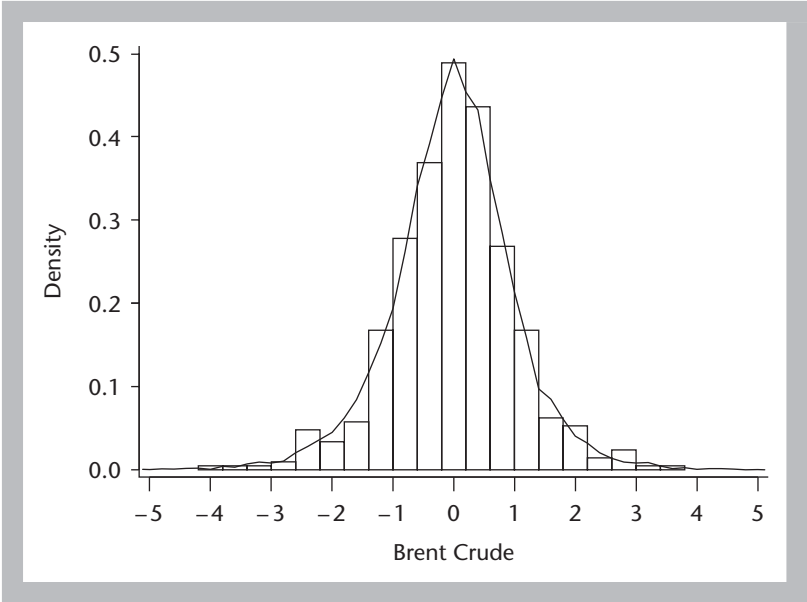
Figure 5.10 Cornish–Fisher adjusted and histogram of daily standardized log price return of Brent Crude (2-month forward)
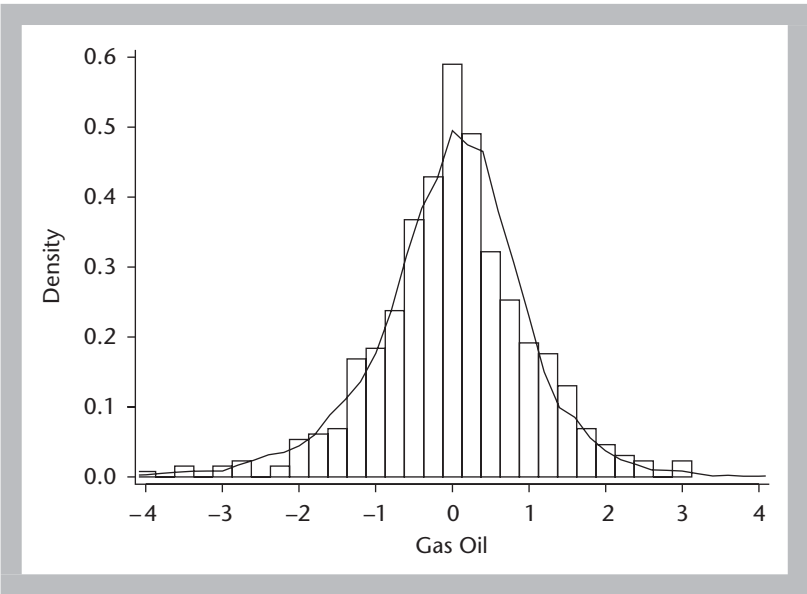
Figure 5.11 Cornish–Fisher adjusted and histogram of daily standardized log price return of Gas Oil (spot)
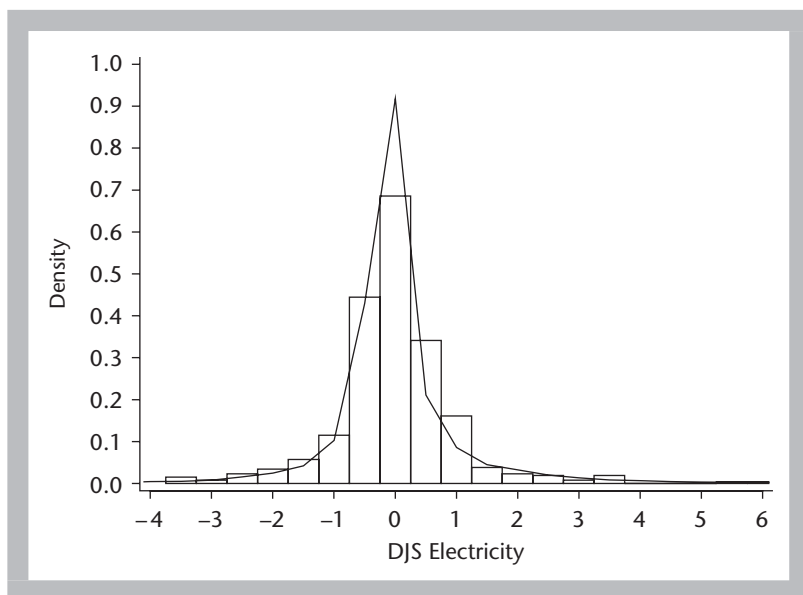
Figure 5.12 Cornish–Fisher adjusted and histogram of daily standardized log price return of the Dow Jones Swiss Electricity price index

5.4.2   Selecting an alternative distribution

Another approach is to select an alternative probability distribution for ε_t. The only question is which distribution should we choose? The best way to address this issue is via the sample descriptive statistics. If we feel O.K of allowing the selected probability distribution to be symmetric, our only concern is choosing a distribution which can exhibit fatter tails than the Normal distribution. Furthermore, since the 2-month Brent forward and Gas Oil are quite similar in terms of their degree of kurtosis and skew we should be able to get away with using the same probability distribution for both of these products. DJS Electricity is a little more challenging; however, to emphasize the limitations we shall continue to include it in our analysis.

5.4.2.1   Logistic distribution

A potential probability distribution for these products is the Logistic distribution. The Logistic distribution is a symmetrical bell-shaped distribution. One of its common applications is as an alternative to the normal distribution when a higher proportion of the observations being modeled are distributed in the tails. This is exactly the situation we face with the energy products discussed so far. Figure 5.13 illustrates the difference between the
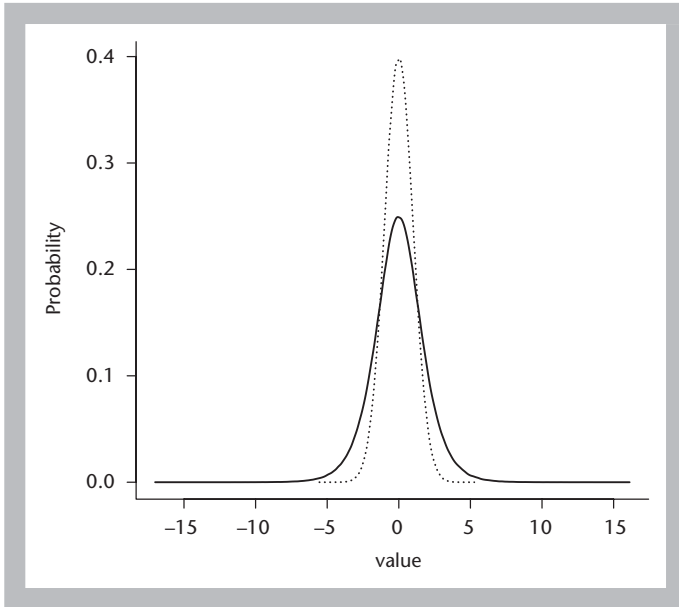
Figure 5.13 Logistic (flatter curve) and normal distributions

Table 5.4 Comparision of the standard normal and standard logistic distributions

Distribution	Kurtosis	5% Percentile	95% Percentile
Logistic	4.2	−1.6234	1.6234
Normal	3.0	−1.6448	1.6448

standard normal and logistic distribution with scale parameter equal to one and location parameter equal to zero. It is immediately obvious that the Logistic distribution contains considerably more probability in the tails – that is it has heavier tails than the normal distribution. Table 5.4 gives a comparison of the standard normal and logistic distribution with a location parameter equal to zero and scale parameter equal to one. The higher value of kurtosis for the Logistic distribution indicates that a higher proportion of the observations are located further away from the mean than in the normal distribution. This is also reflected in the lower values for the 5th and 95th percentiles.

The probability density function of the logistic distribution, with location parameter α and scale parameter β, is given by

$$f(r) = \frac{\exp((r - \alpha)/c)}{c[1 + \exp((r - \alpha)/c)]^2},$$

where c is calculated as

$$c = \sqrt{3}\frac{\beta}{\pi}.$$

Provided we assume $c > 0$ we can estimate α using the sample mean and β using

$$\beta = \sqrt{\left(\frac{3}{\pi^2}S^2\right)}$$

where S is the sample standard deviation.

The logistic distribution has a kurtosis equal to 4.2, which is close to the sample estimates of both the 2-month forward Brent Crude and Gas Oil spot. However, it is considerably less than the sample estimate of kurtosis for Electricity. Figure 5.14 shows the fitted logistic distributions and histograms of the standardized returns. It is immediately apparent that the fitted logistic distribution provides a much better fit than the standard normal distribution. Therefore, for Brent Crude we might wish to specify the model

$$r_t = 0.008 + 0.010365 \times \ddot{\varepsilon}_t,$$

whilst for Gas Oil over the same period we might specify

$$r_t = 0.009 + 0.013452 \times \ddot{\varepsilon}_t,$$

where we assume $\ddot{\varepsilon}_t$ is from the standard logistic distribution. Notice that in this case, given an estimate $\hat{\beta}$, the random error $\ddot{\varepsilon}_t$ is multiplied by $\sqrt{(\pi^2/3)\hat{\beta}^2}$. This is because this quantity is equal to the sample standard deviation. As the sample kurtosis of Electricity is considerably larger than that of the logistic distribution, we would probably elect not to use this distribution to approximate the empirical data.

Selecting an appropriate distribution for empirical data requires a considerable degree of art embedded in the science. Once we have calculated sample estimates of means, standard deviations, skew, and kurtosis, we are in a position to begin to narrow down the choice to a handful of suitable probability distributions. In practice the parameter estimates of many of the potential distributions may not be readily available in a spreadsheet or statistical package and one must then resort to numerical programming. When time is short and the pressure is on I have found the probability distributions listed in the following section to be particularly useful. Alongside the distributions I also give expressions for the parameter estimators. A key
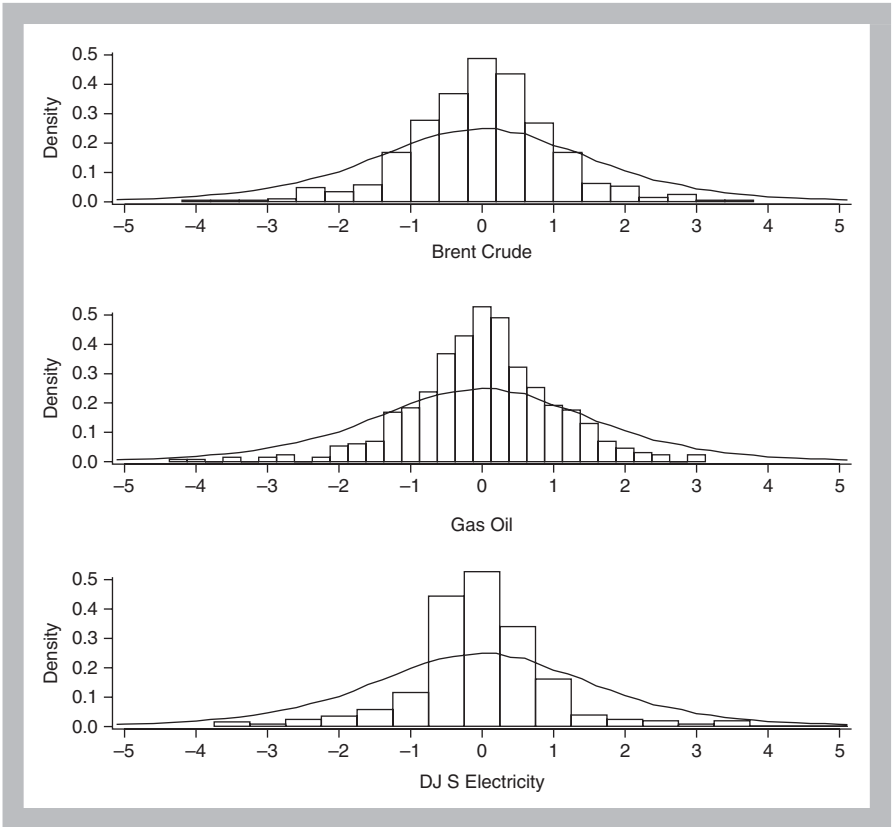
Figure 5.14 Fitted logistic distribution (continuous line) and histogram of daily standardized log price returns

element of my criteria for listing these distributions is that their estimators can be quickly entered into a spreadsheet package, no specialized statistical software is required.

5.4.2.2   *Extreme value (type I) distribution*

Extreme value probability distributions, of which there are three types, are useful for modeling the risk inherent in the tails of a sample. Here we focus on the type I distribution for the smallest extreme value. The probability density function of the extreme value (type I) distribution is given by

$$f(r) = \frac{1}{\beta} \exp\left(\frac{(r-\alpha)}{\beta}\right) \times \exp\left[-\exp\left(\frac{(r-\alpha)}{\beta}\right)\right],$$
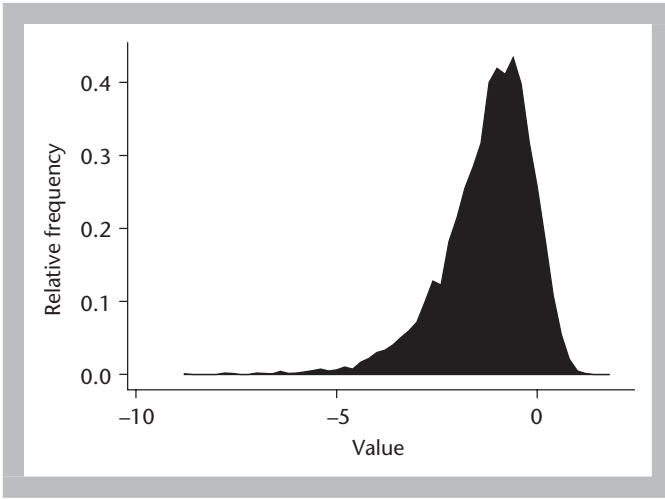
Figure 5.15 Simulated values (10,000) of a type I Extreme Value distribution with location parameter $\alpha = 0$ and scale parameter $\beta = 2$

where α is a location parameter and $\beta > 0$ is a scale parameter. Figure 5.15 shows the results of a simulation of 10,000 observations from this distribution with location parameter equal to 0 and scale parameter equal to 2. The distribution has a skew of around 1.14 and a kurtosis of 5.4. We can calculate the arithmetic mean of the distribution as

$$\text{Mean} = \alpha - 0.57721 \times \beta.$$

The median is

$$\text{Median} = \alpha + \beta \log \log 2.$$

Notice that for this distribution, as opposed to the normal and logistic, the location parameter is not identical to the arithmetic mean. The same is true for the scale parameter. Indeed the variance is calculated as

$$\text{variance} = \frac{\beta^2 \pi^2}{6}.$$

A quick and approximate estimator of the location and scale parameters are given by

1. $\hat{\alpha} = \bar{R} - (0.57221\hat{\beta})$, where $\bar{R}$ is the sample mean.

2. $\hat{\beta} = \left(\dfrac{\sqrt{6}}{\pi}\right) S$, where S is the sample standard deviation.
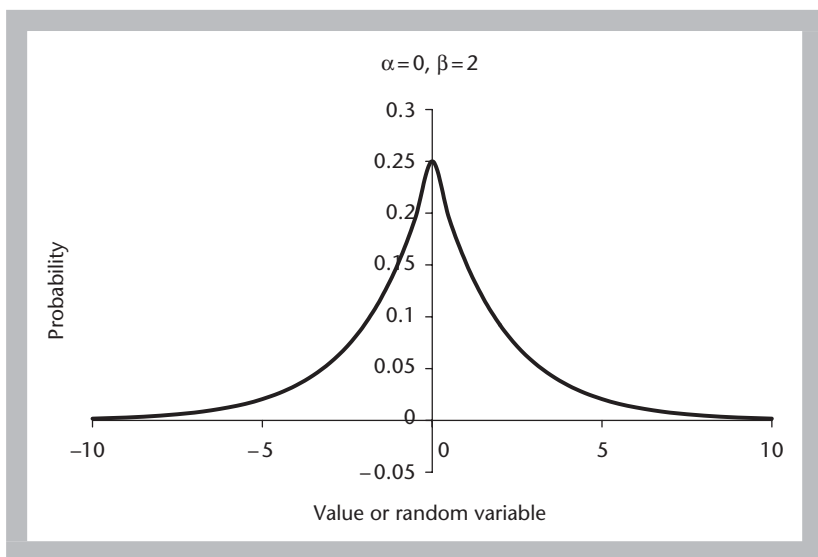
Figure 5.16 Laplace distribution with location parameter $\alpha = 0$ and scale parameter $\beta = 2$

5.4.2.3 Laplace distribution

The probability density function of the Laplace distribution is given by

$$f(r) = \frac{1}{2\beta} \exp\left(-\frac{|r - \alpha|}{\beta}\right),$$

where $-\infty \leq \alpha \leq \infty$ is the location parameter and $\beta > 0$ is the scale para-
meter. The distribution is symmetric with both the mean and median equal
to α. Figure 5.16 illustrates a Laplace distribution with location parameter
equal to 0 and scale parameter equal to 2. The distribution is fat tailed with
a kurtosis of 6. The location parameter can be estimated using the sample
median whilst the scale parameter can be estimated as

$$\hat{\beta} = \sum_{i=1}^{N} \frac{|r_i - \alpha|}{N}.$$

5.4.2.4 Student t-distribution

The Student t-distribution is popular for modeling fat tails of price returns.
The general expression for its probability distribution function is quite
complex. Estimation of the parameters requires optimization via the method
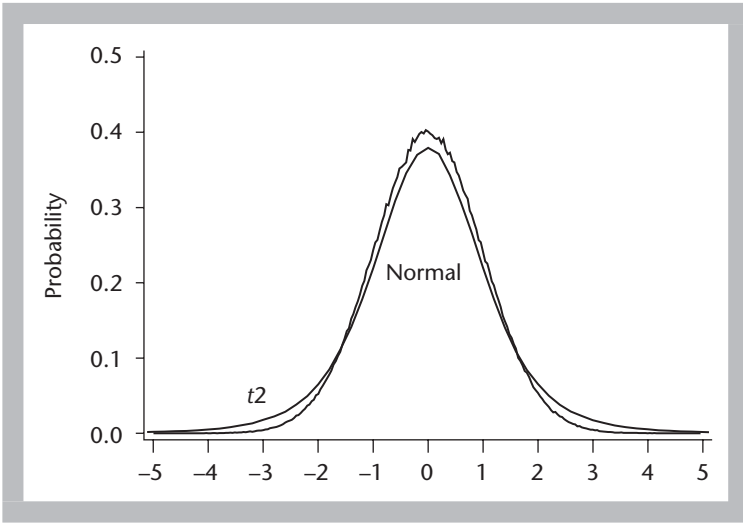
Figure 5.17 Student *t* distribution with 2 degrees of freedom alongside standard normal distribution

of maximum likelihood (see Section 5.6.1). The exact characteristics of the distribution are determined by what are known as degrees of freedom. It just happens to turn out that a Student *t*-distribution with more than around 30 degrees of freedom is approximately normally distributed. The larger the degrees of freedom, the better the approximation to the normal distribution. We have already seen that the normal distribution is not necessarily a good mathematical proxy for energy price returns, hence if we use a Student *t*-distribution we are likely to choose one with low degrees of freedom. It turns out that if we choose 2 degrees of freedom, the Student *t*-distribution has a very simple form given by

$$f(x) = \frac{1}{(2 + r^2)^{3/2}}.$$

In Figure 5.17 the results of simulating 100,000 observations from this distribution are given alongside the standard normal distribution; we see that it is symmetric around zero but has much fatter tails relative to a standard normal distribution.

5.5   BUILDING MODELS USING MIXTURE DISTRIBUTIONS

So far we have developed three models for energy price returns. In

Model 1: $r_t = \mu + \sigma \times \varepsilon_t$, where $\varepsilon_t \sim N(0, 1)$.

This model did not appear adequate for the energy products we discussed. This was primarily due to the apparent failure of the normality assumption surrounding ε_t. The second model used the Cornish–Fisher expansion:

Model 2: $r_t = \mu + \sigma \times \tilde{\varepsilon}_t$, where $\tilde{\varepsilon}_t \sim N_{CFA}(0,1)$ and $N_{CFA}(0,1)$ is the Cornish–Fisher adjustment to the standard normal inverse probability function. In the third model we have

Model 3: $r_t = \mu + \tilde{\sigma} \times \ddot{\varepsilon}_t$, where we assume $\ddot{\varepsilon}_t$ is from some non-normal probability distribution.

In choosing to select a model for energy price returns we might begin with model 1, if we reject it we can try model 2 and if we reject this model, we might then select model 3.

Statistical modeling is often complicated and messy. Outside of the theoretical papers in academic journals, the practicing risk manager demands simplicity and above all interpretability. Since model 1 is as simple as things get it is worth spending a little more time on trying to understand how its basic structure can be altered in order that we might be able to keep the assumption of normality. Keeping things "normal," and the inherent requirement for simplicity, is a sentiment echoed in the wider financial community:

> The biggest problem we now have with the whole evolution of the risk is the fat-tailed problem, which is really creating very large conceptual difficulties. Because as we all know, the assumption of normality enables us to drop off the huge amount of complexity in our equations. Because once you start putting in non-normality assumptions, which is unfortunately what characterises the real world, then these issues become extremely difficult. Dr Alan Greenspan (Chairman of the Board of Governors of the Federal Reserve System of the United States of America, 1997)

To get a feel for a potential source of non-normality consider Figure 5.18 which shows the histogram of price returns for a fictional energy product. Notice the distribution of price returns is skewed and fat tailed. Indeed the coefficient of kurtosis is around 8 and the coefficient of skew is approximately equal to -2. Clearly these price returns cannot be generated from a Normal distribution, or can they? In actual fact they were generated from a mixture of two Normal distributions, the first with a mean of -5.5, the second with a mean of 2. Both distributions have a standard deviation of 1. If we denote the mean of the sample by μ, the mixture equation used to generate the returns was

$$\mu \sim \begin{cases} N(-5,1) \text{ with probability } 0.95 \\ N(2,1) \text{ with probability } 0.05. \end{cases}$$
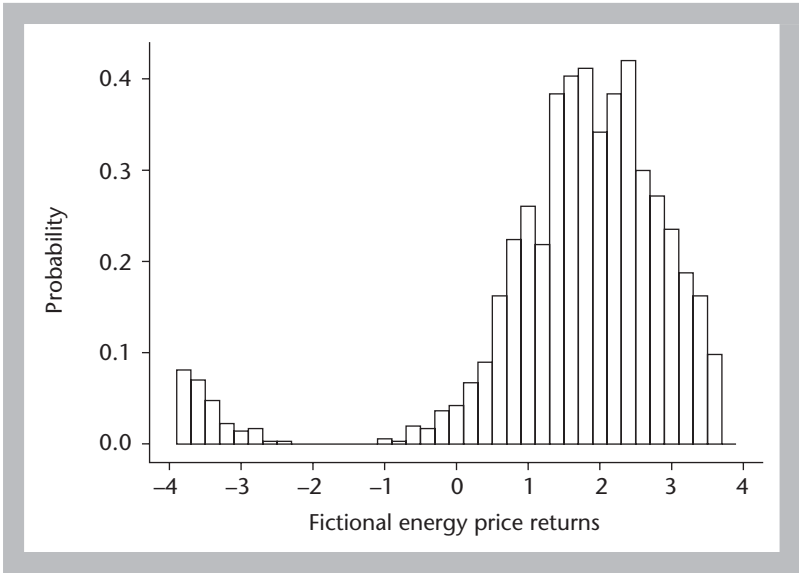
Figure 5.18 Histogram of the simulated price returns of an imaginary energy product

The important point to recognize is that although the price returns were generated by two separate normal distributions, the mixture distribution appears non-normal. It exhibits characteristics which are often evident in energy price returns – skew and kurtosis.

The possibility that energy price returns are generated by a mixture of normal distributions puts the assumption of normality back on the table. Large values for the coefficient of skew and kurtosis in empirical data may simply reflect that returns have been generated by a mixture of normal distributions. Given this possibility, it seems reasonable to adapt model 1 to take this into account

$$r_t = \begin{cases} \mu_1 + \sigma \times \varepsilon_t \text{ with probabilty } \theta, \\ \mu_2 + \sigma \times \varepsilon_t \text{ with probabilty } 1 - \theta, \end{cases} \quad \text{where } \varepsilon_t \sim N(0,1).$$

5.5.1   A simple procedure for estimating mixture distributions

The procedure by which we can estimate the parameters of the above mixture of normals model is as follows.[3]

Step 1: Simulate two large samples, the first from a normal distribution with mean equal to μ_1 and the second with mean equal to μ_2. Both distributions should have a standard deviation equal to 1.
Step 2: Choose a value for the mixing probability θ.

Step 3: Construct the aggregate return series using the mixing probability and both samples.

Step 4: Standardize the aggregate return series so that it has a mean of zero and standard deviation equal to 1.

Step 5: Compare the descriptive shape statistics (skew and kurtosis) of the simulated aggregate distribution to the empirical characteristics of the standardized sample. If they are reasonably close you are done, if not go back to step 1.

The process is iterative and easily implemented in a spreadsheet package or statistical tool such as R. We illustrate the ease with which this can be accomplished in Example 5.1.

Example 5.1   Fitting a mixture of normal distributions to the WEFA Steam Coal index returns

Figure 5.19 shows the histogram of standardized quarterly returns of the WEFA Steam Coal ARA price index over the period March 1996 to December 2003. This index of steam coal prices is commonly used as a benchmark index for derivative trades and also for physical coal transactions. The histogram indicates a moderate degree of asymmetry with a relative large amount of observations in both tails. The coefficient of kurtosis is 5.7 with the coefficient of skew around 0.9. We shall attempt to describe these
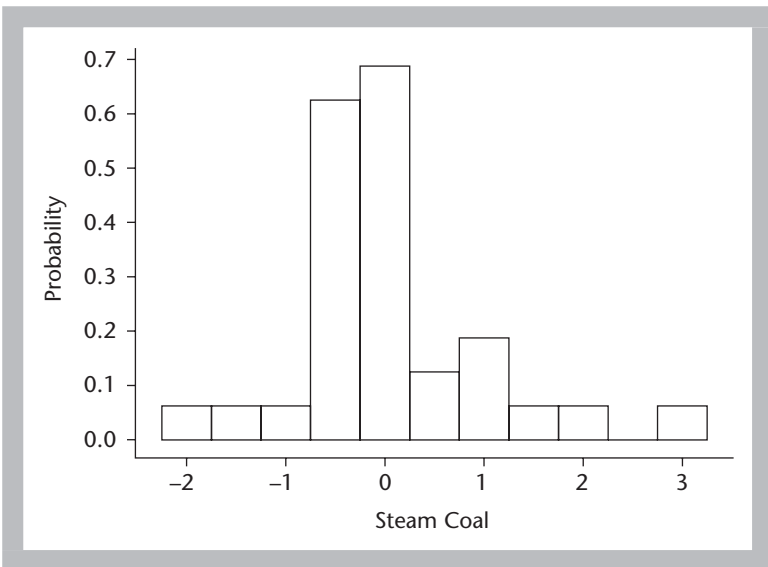
Figure 5.19 Histogram of standardized WEFA Steam Coal ARA quarterly price returns

empirical characteristics using a mixture of two Normal distributions each with standard deviation equal to 1, but with differing means. We begin by simulating two samples of 10,000 observations as follows

Sample $1 \sim N(6,1)$,

Sample $2 \sim N(3,1)$.

We choose as the mixing probability $\theta = 0.9$. The resulting aggregate stand-ardized return series has a skew of 1.4 and kurtosis of 5. Since the skew appears a little high we re-simulated sample 1 with a mean of zero and sample 2 with a mean of -2.5 along with a mixing probability of 0.995. In this case the skew for the aggregate distribution is around 0.6 with a kurtosis of 5.8. Since this is close to the values observed in the empirical data, our final model for the standardized returns takes the form

$$r_t = \begin{cases} \varepsilon_t \text{ with probability } 0.995, \\ -2.5 + \varepsilon_t \text{ with probability } 0.005, \end{cases} \quad \text{where } \varepsilon_t \sim N(0,1).$$

The fitted mixture probability distribution alongside the standardized returns are shown in Figure 5.20. Visual inspection reveals the distribution fits the data fairly well. However, the fit, especially the right tail, can be considerably improved by further experimentation.
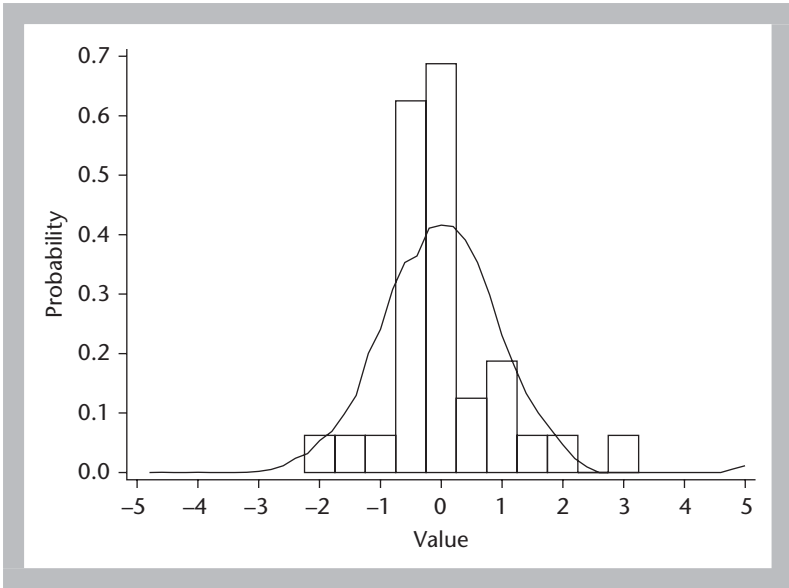
Figure 5.20 Fitted normal–normal mixture distribution and histogram of WEFA Steam Coal ARA quarterly price returns

5.5.2   A generic model for mixture distributions

We can obtain more flexibility in our fitting if we generalize the mixture model to the situation where energy price returns have been generated by k different normal distributions. In this case our model would take the form

$$r_t = \begin{cases} \mu_1 + \sigma \times \varepsilon_t^1 \text{ with probability } \theta_1, \\ \mu_2 + \sigma \times \varepsilon_t^2 \text{ with probability } \theta_2, \\ \quad\quad\quad . \\ \quad\quad\quad . \\ \quad\quad\quad . \\ \mu_{k-1} + \sigma \times \varepsilon_t^{k-1} \text{ with probability } \theta_{k-1}, \\ \mu_k + \sigma \times \varepsilon_t^k \text{ with probability } \theta_k, \end{cases} \quad \text{where } \varepsilon_t \sim N(0,1)$$

Of course we could go even further and remove the assumption of normality of the εs. If we denote an arbitrary probability distribution $\Phi(\phi_1, ..., \phi_n)$, where $\{\phi_1, ..., \phi_n\}$ are its parameters we could specify an even more general model of the form:

$$r_t = \begin{cases} \mu_1 + \sigma \times \varepsilon_t^1 \text{ with probability } \theta_1, \\ \mu_2 + \sigma \times \varepsilon_t^2 \text{ with probability } \theta_2, \\ \quad\quad\quad . \\ \quad\quad\quad . \\ \quad\quad\quad . \\ \mu_{k-1} + \sigma \times \varepsilon_t^{k-1} \text{ with probability } \theta_{k-1}, \\ \mu_k + \sigma \times \varepsilon_t^k \text{ with probability } \theta_k, \end{cases} \quad \text{where } \varepsilon_t^i \sim \Phi^i(\phi_1, ..., \phi_n).$$

In this model $\Phi^i(.)$ and $\Phi^j(.)(i \neq j)$ may be completely different probability distributions. In practice fitting the above model will prove complicated, especially as k increases. It can be done as shown in Example 5.2, but involves estimating both the θs and the parameters of k possibility different probability distributions.

Example 5.2   Fitting a normal–logistic mixture distribution to the WEFA Steam Coal index returns

Figure 5.21 shows the fitted mixture distribution derived from a mixture of a standard Logistic distribution (with scale parameter equal to 1 and location parameter equal to 0) and a normal distribution (with mean equal to -0.06 and standard deviation equal to 1). The mixing parameter θ was set equal to 0.5. These values were obtained by following the procedure described previously. In this case the first sample was simulated from a standard logistic distribution and the second sample from a normal distribution. The fitted
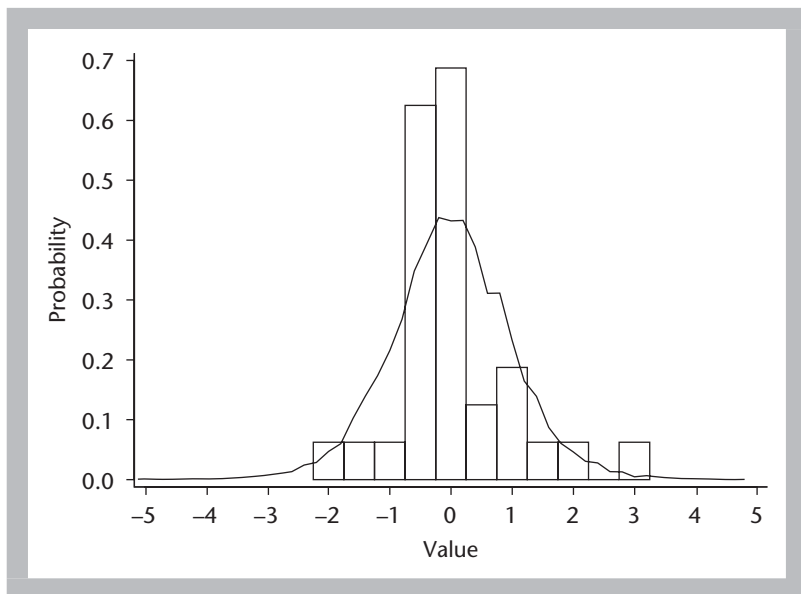
Figure 5.21 Fitted logistic–normal mixture distribution and histogram of WEFA Steam Coal ARA quarterly price returns

distribution is marginally more peaked than the fitted Normal–Normal mixture distribution of Figure 5.20.

Ideally in fitting mixture distributions we would like to keep k small say equal to 2 or 3. A useful starting point is to try to capture the characteristics of the empirical data with two normal distributions, failing that try three normal distributions and so on. If this fails then it might be feasible to try a normal and say logistic or student t-distribution. This is very much an adhoc approach and it should be noted that as k increases we rapidly depart from our core objective of parsimony.

5.6   GENERAL APPROACHES FOR ESTIMATING PARAMETERS

All probability distributions, including the now familiar Normal and Logistic, are totally determined by their parameters. In order to obtain a suitable probability model for our energy price return data we need to obtain an estimate of the parameters of the postulated probability model. By now you should be relatively comfortable in specifying and assessing the validity of a probability model. The question is how can we obtain parameter estimates for those probability distributions not discussed so far in this chapter? For the Normal distribution we were fortunate in being able to use the sample estimator of the mean and variance as these corresponded

to the parameters of the probability distribution. For other distributions this will not be the case and we will need to obtain a suitable estimator. In practice there are three general approaches we can use to obtain a suitable estimator. The first approach involves plotting the sample data against the cumulative probability function of the postulated probability distribution on special graphical paper. The value of the estimated parameters is then read off the graphical plot. This approach is useful for quick approximations. In the second approach a system of equations equal to the number of parameters to be estimated is solved to yield the estimators. Common techniques in this category are the Method of Moments, Probability Weighted Methods, Order Statistics and Percentile Matching. The third approach includes optimization methods which yield estimators by maximizing or minimizing some function of the data. Typical techniques in this category are the methods of maximum likelihood and ordinary least squares. In theory any of the above methods could be used to obtain parameter estimates. However, two techniques namely, the method of maximum likelihood and ordinary least squares, have dominated because they produce estimators that have good statistical properties and can be easily programmed into statistical packages.[4]

5.6.1   The method of maximum likelihood

The method of maximum likelihood is one way by which we can obtain an estimator. We already know that an estimator is essentially a function of the data. For example the sample mean is an estimator of the population mean – it is a simple function of the sample data. The same can be said for the sample variance, skew and so on. Sample estimators of the parameters of probability distributions are also some functions of the sample data and hence their specific values will be dependent on the observed values in the sample. The method of maximum likelihood makes good use of this property. It chooses the most likely values for the parameters given the observed data. This is achieved by maximizing what is known as the likelihood equation of the postulated probability model. The likelihood function of a sample is the probability of observing the actual sample given the postulated probability model under the assumption that the observations are all from the same underlying probability distribution with given (unknown) parameters and statistically independent of each other. This assumption is often shorted to independent identically distributed.

Given a sample of observations on the price return of some energy product $\{r_1, ..., r_N\}$, and a postulated probability density, $f(r)$, which depends on k parameters $\{\Theta_1, ..., \Theta_k\}$ whose values we wish to estimate; The likelihood

equation is given by

$$L(\Theta_1, ..., \Theta_k | r_1, ..., r_N) = \prod_{i=1}^{N} f(r_i)$$

The likelihood equation is interpreted as a function of the parameters for a given sample $\{r_1, ..., r_N\}$. The objective is to find the values of the parameters $\{\Theta_1, ..., \Theta_k\}$ that make the observed values of $\{r_1, ..., r_N\}$ most likely. The most likely values of the parameters are those that maximize the likelihood function. In practice it is often more convenient to work with the log likelihood

$$\log L(\Theta_1, ..., \Theta_k | r_1, ..., r_N) = \sum_{i=1}^{N} f(r_i).$$

5.6.1.1 Analytically solving the likelihood function

Once we have specified a likelihood function, we seek to find those values of the parameters that maximize it. One approach, illustrated in Example 5.3, is to take partial derivatives with respect to each of the parameters, set these derivatives equal to zero and solve the resulting set of equations for the unknown parameters.

Example 5.3   Likelihood and log likelihood function of an exponentially distributed sample

Suppose a energy risk manager is only interested in the absolute value of negative returns of a particular product. She collects a sample $\{r_1, ..., r_N\}$ which she assumes is independent identically distributed with probability density function:

$$f(r_i) = \lambda \exp(-\lambda r_i)$$

This is the probability density of the exponential distribution. The likelihood function for the sample is

$$L(\lambda | r_1, ..., r_N) = \prod_{i=1}^{N} f(r_i) = \prod_{i=1}^{N} \lambda \exp(-\lambda r_i).$$

Taking the natural logarithm yields the log likelihood

$$\log L(\lambda | r_1, ..., r_N) = N \log(\lambda) - \lambda \sum_{i=1}^{N} r_i.$$

The partial derivatives of the log likelihood function are given by

$$\frac{\partial \log L(\lambda | r_1, ..., r_N)}{\partial \lambda} = \frac{N}{\lambda} - \sum_{i=1}^{N} r_i = 0$$

Solving for λ, the maximum likelihood estimator is therefore

$$\hat{\lambda} = \frac{1}{(1/N) \sum_{i=1}^{N} r_i}.$$

Recall the previous comment that an estimator is a function of the sample data. Take a close look at the above equation – does it seem familiar? It should be, for it is simply 1/sample mean. Of course not all estimators are this easy to interpret. But it does serve to reinforce the point that an estimator is a function of the sample data. We can use the same analytical technique to obtain the maximum likelihood parameters for many probability distributions; Example 5.4 illustrates this for the normal distribution.

Example 5.4   Analytical solution to the log likelihood function of an normally distributed sample

The normal likelihood function is given by

$$L(\mu \sigma^2 | r_1, ..., r_N) = \prod_{i=1}^{N} f(r_i) = \prod_{i=1}^{N} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(r_i - \mu)^2\right) \right\}$$

$$= (2\pi\sigma^2)^{-(1/2N)} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{N} \frac{(r_i - \mu)^2}{\sigma^2} \right\}.$$

The log likelihood function can be written as

$$L(\mu \sigma^2 | r_1, ..., r_N) = -\frac{1}{2} N \log 2\pi\sigma^2 - \frac{1/2 \sum_{i=1}^{N} (r_i - \mu)^2}{\sigma^2}.$$

The analytical solution involves obtaining estimates that maximize the likelihood function. As we have already seen this involves taking partial

derivatives with respect to each of the parameters. These derivatives are then set equal to zero and the estimators arise from solving the resulting set of equations. The first derivative of the log likelihood function with respect to μ is

$$\frac{\partial \log L(\mu, \sigma^2 | r_1, ..., r_N)}{\partial \mu} = \sum_{i=1}^{N} \frac{(r_i - \mu)}{\sigma^2}.$$

Setting equal to zero and solving

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} r_i$$

The first derivative of the log likelihood function with respect to σ is given by

$$\frac{\partial \log L(\mu, \sigma^2 | r_1, ..., r_N)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \sum_{i=1}^{N} \frac{(r_i - \mu)^2}{2\sigma^4}.$$

Setting equal to zero and solving

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (r_i - \hat{\mu})^2.$$

Therefore the maximum likelihood estimators of μ and σ are

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} r_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (r_i - \hat{\mu})^2.$$

In the above example the estimator $\hat{\mu}$ is simply the arithmetic mean. This is to be expected, as we previously saw the parameters of the Normal distribution can be estimated by the sample mean and sample standard deviation. The estimator for the variance feels familiar yet somehow different. In fact it is very similar to the sample estimator of the variance

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (r_i - \hat{\mu})^2.$$

The only difference is that the maximum likelihood estimator divides by N rather than $N-1$. The maximum likelihood estimator is thus a biased estimator of variance. It should actually divide by $N-1$ and not N. Whilst this

is not particularly important in large samples it can be a potential problem in small samples. To see this consider the standardized quarterly returns of the WEFA Steam Coal ARA price index over the period March 1996 to December 2003. Since the returns are standardized they will have a mean of 0 and a variance of 1. The maximum likelihood estimator of the mean corresponds to the sample estimate, in this case 0. However, the maximum likelihood estimate of the variance is 0.969, slightly less than the sample variance of 1. Therefore in practice, when estimating the parameters of the normal distribution, we should use the sample mean and sample variance as estimators for the parameters of the normal distribution rather than the maximum likelihood estimators. A key point to recognize is that maximum likelihood estimators are not necessarily the optimum estimators.

5.6.1.2   Numerically solving the likelihood function

Analytically solving the log likelihood function for the parameters of interest is very tedious and frequently impossible. Fortunately, most statistical software have an optimisation function. For example the freely available statistical package R has a function `optimise()`.[5] This function takes as arguments:

1. the log likelihood (or other function you wish to maximise),

2. a range over which to numerically optimize the function,

3. a tolerance threshold for optimizing the function,

4. the sample data and an argument TRUE or FALSE for maximization and minimization respectively.

Spreadsheet packages such as Microsoft Excel also offer optimization tools. Excel has the solver tool which can be used to maximize or minimize a function. To illustrate the use of an optimization function we return to the statistical package R. We begin by simulating 100 observations from an exponential distribution with parameter $\lambda = 10$

```
> sample =rexp(100,10)
```

We already know that the an estimator of the parameter λ is 1/sample mean. The sample mean can be calculated using

```
> mean(sample)
[1]  0.10212
> 1/mean(sample)
[1]  9.792.
```

Hence we estimate $\hat{\lambda} = 9.792$, which although not exactly equal to 10, is close enough given the sample size. Let us attempt to estimate the parameter λ using the method of maximum likelihood. We begin by specifying the log likelihood function

$$\log L(\lambda|r_1,...,r_N) = N\log(\lambda) - \lambda \sum_{i=1}^{N} r_i.$$

In the statistical package R we can do this using the following command to define the log likelihood function

```
> log.likelihood<-function(a,r) length(r)*log(a)-a *sum(r).
```

Now we can use the function `optimize()` to maximize the function `log.likelihood`:

```
>estimate=optimise(log.likelihood,c(0,20),tol=.0000001,
r=sample,maximum=TRUE)
> estimate
$maximum
[1] 9.792
$objective
[1] 128.16
```

The output informs us that the log likelihood is maximized at a value of 128.16 with the estimate $\hat{\lambda}$ equal to 9.792.

5.6.1.3 Maximum likelihood in a nutshell

Behind the equations and numerical procedures required to obtain maximum likelihood estimates is a very simple notion. We choose as our estimates those values of the parameters of a probability distribution which maximise the joint probability of them coming from the specific sample. Thus the estimate of $\hat{\lambda} = 9.792$ given above, implies, given the sample, that 9.792 is the most likely value for the parameter of the underlying probability distribution. It is more likely than a value of 8.98 and more likely than a value of 10.79. It is that value which has the maximum probability of coming from the given sample. Of course, since we are dealing with samples, we would expect the value of the estimate to be close to, but not exactly match, the true population parameter value of 10. We could if we wished to construct a confidence interval around the estimate in an attempt to measure this uncertainty.

5.6.2 The method of Ordinary Least Squares

For much of the time we will be content to use the method of maximum likelihood. However on occasions, as we will see in Chapter 8, we will choose to use the method of ordinary least squares (OLS) to arrive at an estimate. To illustrate this method consider the function

$$Y_t = \beta_1 + \beta_2 X_{1t} + \beta_3 X_{3t} + \cdots + \beta_k X_{kt},$$

where $\{\beta_1, ..., \beta_k\}$ are unknown model parameters for the distinct variables $\{X_1, ..., X_k\}$. If we define the observed model error as

$$\varepsilon_t = Y_t - (\beta_1 + \beta_2 X_{1t} + \beta_3 X_{3t} + \cdots + \beta_k X_{kt}).$$

The method of ordinary least squares chooses the parameter values in such a way as to minimize the square of the observed model errors

$$\sum_{t=1}^{N} \epsilon_t^2.$$

The underlying idea can be more easily appreciated in matrix notation. Let **Y** be the vector whose elements are $\{y_1, ..., y_n\}$, **X** be a matrix whose columns contain the K variables $\{X_1, ..., X_k\}$ and β a vector containing the unknown coefficients $\{\beta_1, \beta_2, ..., \beta_k\}$. The method of OLS calculates estimates of the coefficient vector denoted by $\hat{\beta}$ so that the vector $\mathbf{X}\hat{\beta}$ minimizes the sum of squared elements from the deviation

$$\epsilon = \mathbf{Y} - \mathbf{X}\hat{\beta}.$$

The solution is given by

$$\hat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

5.6.2.1 *Application of OLS in the R statistical package*

As with the method of maximum likelihood, OLS is contained in most statistical and spreadsheet packages. We shall illustrate its basic calculation using the matrix manipulation capabilities of the R statistical package. Consider the data on two variables Y and X where

```
Y = {12.66,19.87,1.19,2.97,13.48,2.29,6.77},
```

and

$$X = \{7.97, 11.25, 3.26, 3.12, 8.33, 2.81, 4.24\}.$$

This information can be entered directly into R as

```
> Y=as.matrix(c(12.66,19.87,1.19,2.97,13.48,2.29,6.77)),
> X=as.matrix(cbind(1,c(7.97,11.25,3.26,3.12,8.33,2.81,
  4.24))).
```

Checking to ensure we have entered the data correctly we see

```
> Y
        [,1]
[1,] 12.66
[2,] 19.87
[3,]  1.19
[4,]  2.97
[5,] 13.48
[6,]  2.29
[7,]  6.77
> X
      [,1]   [,2]
[1,]     1  7.97
[2,]     1 11.25
[3,]     1  3.26
[4,]     1  3.12
[5,]     1  8.33
[6,]     1  2.81
[7,]     1  4.24
```

Now we can use the function t() to obtain the transpose of X, and we store it in the variable XT

```
> XT <- t(X)
> XT
      [,1]   [,2] [,3] [,4] [,5] [,6] [,7]
[1,] 1.00   1.00 1.00 1.00 1.00 1.00 1.00
[2,] 7.97 11.25 3.26 3.12 8.33 2.81 4.24
```

We also need a variable to represent $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, we call this variable XTX. Multiplication of the vectors is carried out using the operator "%*%"

```
> XTX<-t(X)%*% X
> XTX
       [,1]    [,2]
[1,]  7.00  40.980
[2,] 40.98 305.708.
```

Next we use `solve()` to obtain the inverse $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ and the result is stored in the variable XTXINV

```
> XTXINV<-solve(XTX)
> XTXINV
              [,1]          [,2]
[1,]  0.66372323 -0.08897176
[2,] -0.08897176  0.01519771
```

To obtain $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ we again use the matrix multiplication operator "%*%" and store the result in the variable XTY

```
> XTY <-XT %*% Y
> XTY
          [,1]
[1,]  59.2300
[2,] 485.0116.
```

Finally we can calculate the expression $\hat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$

```
> Beta.Est = XTXINV %*% XTY
> round(Beta.Est,3)
        [,1]
[1,] -3.840
[2,]  2.101.
```

We see that β_1, the intercept term, is equal to -3.84 and the coefficient β_2 is equal to 2.101. Fortunately most statistical and spreadsheet packages hide the details on these calculations, we need not dwell on them any further, although if asked, you now know how they work.

5.6.2.2 Estimating the parameters of a probability distribution using OLS

Recently, autoregressive conditional duration models have found popularity as a statistical model for the analysis of data that does not arrive in equal time intervals such as energy price data and sales data on products that are tracked electronically. In contrast to more traditional fixed interval analysis, the model treats the time between observation arrivals as a stochastic time varying process. In other words, the times the observations occur, termed "arrival times," are treated as random so that the time duration between observations is a random variable. In practice, when looking at the tick by tick data of energy prices, one observes duration clustering – that is, a short duration tends to follow another short duration while a long duration tends to follow another long duration. A simple model of duration is given by:

Duration = expected duration × random error.

Autoregressive conditional duration models assume a specific form for the "expected duration" term. As our interest in this chapter lies in using the method of OLS to estimate the parameters of a probability distribution, for brevity's sake we shall assume that the expected duration is a constant. This leaves us with the random error term, which can be fitted by log normal, log logistic and Weibull distributions. As an illustration, we consider fitting a Weibull distribution for the term "random error." The probability density function of a Weibull distributed random variable X is given by

$$f(x) = \left(\frac{\beta x^{\beta-1}}{\alpha^\beta} \right) \exp(-(x/\alpha))^\beta.$$

β is a shape parameter and α a scale parameter. An estimate of α can be obtained as follows

$$\hat{\alpha} = \exp \left(\tilde{Y} - \left(\frac{\tilde{X}}{\hat{\beta}} \right) \right),$$

where

$$\tilde{X} = \frac{1}{N} \sum_{i=1}^{N} \ln \left\{ \ln \left[\frac{1}{(1 - i/(N + 1))} \right] \right\} \quad \text{and} \quad \tilde{Y} = \frac{1}{N} \sum_{i=1}^{N} \ln x_i.$$

An estimate of β is obtained by first calculating the mean rank, denoted by P_i

$$P_i = \frac{(\text{Rank}(x_i) - 0.3)}{(N - 0.4)} \quad \text{and} \quad P_n = \frac{N}{(N + \delta)}$$

where δ is a small positive number, that is, 1E-20. Second, calculate the transformed rank, denoted by T_i

$$T_i = \left\{ \ln \left(\frac{1}{1 - P_i} \right) \right\}.$$

Third, take the natural log the xs

$$y_i = \ln(x_i)$$

Finally, calculate as the OLS coefficient between the T_is and the y_is to arrive at an estimate of β.

Example 5.5   OLS estimation of the parameters of the Weibull distribution for duration data

To illustrate this procedure consider the duration data shown in Table 5.5. The first column shows the original duration data, sorted by rank. The second column gives the rank of each observation. For convenience the observations are presented in size order and the rank increases from 1 (first observation) to 100 (last observation). The third column gives the values for P_i. The forth column presents the values of the natural logarithm of the original data given in the first column. Calculating the OLS coefficient between T_i and y_i yields an estimate of β as $\hat{\beta} = 2.531$. Since $\tilde{X} = -0.5600$ and $\tilde{Y} = 1.6169$, the estimate of α is given by

$$\hat{\alpha} = \exp \left(1.6169 - \left(\frac{-0.56}{2.531} \right) \right) = 6.285.$$

The fitted values alongside the histogram of the original data are given in Figure 5.22. The Weibull distribution appears to be reasonably adequate for this data.

5.7   SUMMARY

Working with probability distributions in applied modeling will involve estimation of the distributions parameters alongside assessing model adequacy. We focused initially in this chapter on the more common probability distributions and provided estimation techniques which can be quickly implemented and have been used successfully across many fields. If the price return is normally distributed the parameters σ and μ summarize the entire distribution. These parameters can be estimated from the sample standard

Table 5.5 Duration data and OLS calculation of Weibull parameters

Original data	Rank	P_i	T_i	y_i
0.9494	1	0.00703	−4.95431	−0.05193
0.9593	2	0.01707	−4.06194	−0.04155
1.1419	3	0.02711	−3.59420	0.13269
1.1998	4	0.03715	−3.27396	0.18215
1.9392	5	0.04719	−3.02953	0.66228
2.1377	6	0.05723	−2.83137	0.75973
2.204	7	0.06727	−2.66444	0.79027
2.7705	8	0.07731	−2.51998	1.01903
2.7739	9	0.08735	−2.39249	1.02025
2.8296	10	0.09739	−2.27824	1.04014
2.8667	11	0.10743	−2.17463	1.05316
2.8942	12	0.11747	−2.07974	1.06271
3.0622	13	0.12751	−1.99213	1.11913
3.0754	14	0.13755	−1.91069	1.12343
3.3185	15	0.15000	−1.81696	1.19951
3.3996	16	0.16000	−1.74667	1.22366
3.4619	17	0.17000	−1.68024	1.24182
3.6036	18	0.18000	−1.61721	1.28193
3.6414	19	0.19000	−1.55722	1.29237
3.7149	20	0.20000	−1.49994	1.31235
3.871	21	0.21000	−1.44510	1.35351
4.0675	22	0.22000	−1.39247	1.40303
4.0751	23	0.23000	−1.34184	1.40490
4.2768	24	0.24000	−1.29303	1.45321
4.3357	25	0.25000	−1.24590	1.46688
4.3932	26	0.26000	−1.20030	1.48006
4.3993	27	0.27000	−1.15610	1.48145
4.4104	28	0.28000	−1.11321	1.48397
4.4461	29	0.29000	−1.07151	1.49203
4.4568	30	0.30000	−1.03093	1.49443
4.4748	31	0.31000	−0.99138	1.49846
4.4759	32	0.32000	−0.95279	1.49871
4.5902	33	0.33000	−0.91510	1.52392

Table 5.5 Continued

Original data	Rank	P_i	T_i	y_i
4.5926	34	0.34000	−0.87824	1.52445
4.6815	35	0.35000	−0.84215	1.54362
4.7687	36	0.36000	−0.80679	1.56207
4.8997	37	0.37000	−0.77211	1.58917
4.9034	38	0.38000	−0.73807	1.58993
4.9088	39	0.39000	−0.70462	1.59103
4.9823	40	0.40000	−0.67173	1.60589
4.991	41	0.41000	−0.63935	1.60764
4.9999	42	0.42000	−0.60747	1.60942
5.0009	43	0.43000	−0.57604	1.60962
5.047	44	0.44000	−0.54504	1.61879
5.0808	45	0.45000	−0.51444	1.62547
5.1339	46	0.46000	−0.48421	1.63587
5.1607	47	0.47000	−0.45432	1.64107
5.2612	48	0.48000	−0.42476	1.66036
5.2819	49	0.49000	−0.39550	1.66429
5.2973	50	0.50000	−0.36651	1.66720
5.305	51	0.51000	−0.33778	1.66865
5.3092	52	0.52000	−0.30929	1.66944
5.337	53	0.53000	−0.28101	1.67466
5.3528	54	0.54000	−0.25292	1.67762
5.4369	55	0.55000	−0.22501	1.69321
5.6229	56	0.56000	−0.19726	1.72685
5.6406	57	0.57000	−0.16964	1.72999
5.6438	58	0.58000	−0.14214	1.73056
5.7621	59	0.59000	−0.11474	1.75130
5.8007	60	0.60000	−0.08742	1.75798
5.8046	61	0.61000	−0.06017	1.75865
5.8232	62	0.62000	−0.03295	1.76185
5.9148	63	0.63000	−0.00576	1.77746
5.9264	64	0.64000	0.02142	1.77942
6.022	65	0.65000	0.04862	1.79542
6.0345	66	0.66000	0.07586	1.79749
6.126	67	0.67000	0.10315	1.81254

Table 5.5 Continued

Original data	Rank	P_i	T_i	y_i
6.3373	68	0.68000	0.13053	1.84645
6.3614	69	0.69000	0.15801	1.85025
6.4845	70	0.70000	0.18563	1.86941
6.6287	71	0.71000	0.21340	1.89141
6.6699	72	0.72000	0.24135	1.89760
6.7237	73	0.73000	0.26952	1.90564
6.7501	74	0.74000	0.29793	1.90956
6.8693	75	0.75000	0.32663	1.92706
7.0485	76	0.76000	0.35566	1.95281
7.0593	77	0.77000	0.38504	1.95435
7.1493	78	0.78000	0.41484	1.96701
7.2452	79	0.79000	0.44510	1.98034
7.324	80	0.80000	0.47588	1.99116
7.3418	81	0.81000	0.50726	1.99358
7.4368	82	0.82000	0.53930	2.00644
7.8928	83	0.83000	0.57208	2.06595
7.9234	84	0.84000	0.60573	2.06982
8.0839	85	0.85000	0.64034	2.08987
8.1803	86	0.86000	0.67606	2.10173
8.2239	87	0.87000	0.71306	2.10704
8.4221	88	0.88000	0.75154	2.13086
8.5585	89	0.89000	0.79176	2.14692
8.5817	90	0.90000	0.83403	2.14963
8.6007	91	0.91000	0.87877	2.15184
8.6661	92	0.92000	0.92653	2.15942
9.0781	93	0.93000	0.97805	2.20586
9.4551	94	0.94000	1.03440	2.24655
9.6145	95	0.95000	1.09719	2.26327
9.8316	96	0.96000	1.16903	2.28560
9.8631	97	0.97000	1.25463	2.28880
9.9554	98	0.98000	1.36405	2.29812
10.073	99	0.99000	1.52718	2.30986
10.0924	100	1.00000	1.00000	2.31178
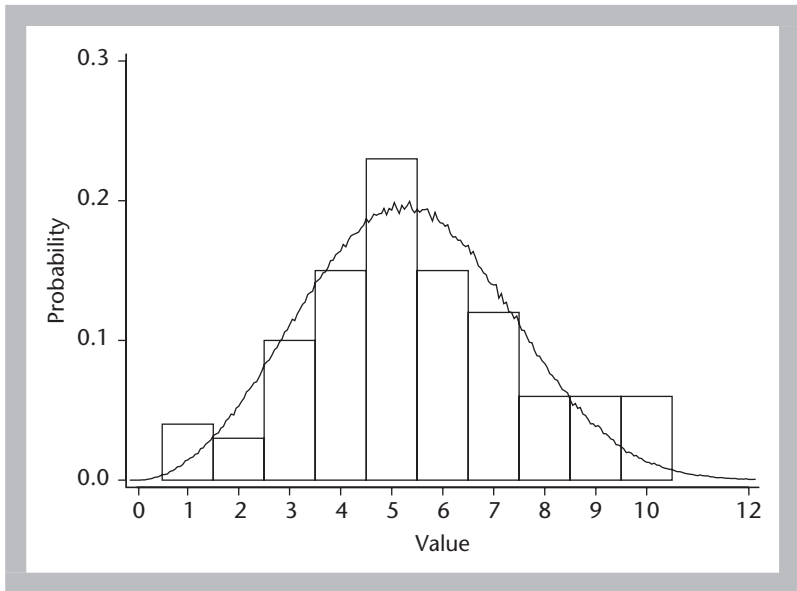
Figure 5.22 Fitted Weibull distributions and histogram of duration data

deviation and sample mean respectively. In fitting other probability models to sample data, the general approach is first to select a specific class of probability distributions which match the sample descriptive statistics such as skew and kurtosis; and then find the values for the parameters of the proposed probability distribution(s) that best match the observed data.

Parameters of probability distributions can also be estimated via methods such as OLS or the method of maximum likelihood. The method of maximum likelihood is one of the most widely used estimation procedures. In this approach the estimates are those values of the postulated probability distribution parameters that make the observed sample data most probable. The procedure can be summarized as follows:

Step 1: Collect sample data.

Step 2: Postulate a probability model for the data.

Step 3: Assume observations are independently distributed from our postulated probability model and then write out the log likelihood function.

Step 4: Maximize the log likelihood function. The maximised values will yield an estimate of the parameters of the postulated probability distribution.

5.8   FURTHER READING

Details of the distributions given in this chapter and many others can be found in Gumbel (1954), Aitchison and Brown (1957), Ascher (1981), Hahn and Shapiro (1967),

Johnson, *et al*. (1994), and Johnson, *et al*. (1995). Lewis (2003) gives a comprehensive list of useful probability distributions alongside easy to implement formulas for parameter estimation. In addition Lewis (2004) provides a CD containing spreadsheets and Excel macros (Visual Basic) that estimate some of the more popular probability models. The formulas can be easily transcribed to other computing packages. In the final analysis which specific approach one eventually adopts when fitting a probability distribution to energy price returns is part of the art of applied modeling and will depend heavily on the particular context.

Aitchison, J. and Brown, J. A. C. (1957) The Log-normal distribution, Cambridge University Press, New York and London.

Ascher, H. (1981) "Weibull Distribution vs Weibull Process," Proceedings, *Annual Reliability and Maintainability Symposium*, pp. 426–31.

Deb, P. and Sefton, M. (1996) The distribution of a Lagrange multiplier test of normality, *Economics Letters*, 51, 123–30.

Gumbel, E. J. (1954) *Statistical Theory of Extreme Values and Some Practical Applications*, National Bureau of Standards Applied Mathematics Series 33, US Government Printing Office, Washington, DC.

Hahn, G. J. and Shapiro, S. S. (1967) *Statistical Models in Engineering*, John Wiley & Sons, Inc., New York.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994) *Continuous Univariate Distributions* Volume 1, 2nd edn, John Wiley & Sons, Inc., New York.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995), *Continuous Univariate Distributions* Volume 2, 2nd edn, John Wiley & Sons, Inc., New York.

Lewis, Nigel Da Costa (2003) *Market Risk Modeling: Applied Statistical Methods for Practitioners*. Risk Books. London.

Lewis, Nigel Da Costa (2004) *Operational Risk with Excel and VBA: Applied Statistical Methods for Risk Management*. John Wiley & Sons, Inc., New York.

Urzúa, C. M. (1996) "On the correct use of omnibus tests for normality," *Economics Letters*, 53, 247–51 (corrigendum, 1997, 54: 301).

5.9   REVIEW QUESTIONS

1 What are the strengths and weaknesses of the simple model for energy returns outlined in this chapter?

2 How might you adapt the simple model to make it a better proxy of actual energy returns?

3 Explain the difference between an estimator and an estimate.

4 Describe the simple procedure for fitting a mixture of distributions to price returns. What are the advantages and pitfalls of this approach?

5 What is a likelihood function? What assumptions do you need to make before specifying it?

6 Write out the likelihood function for the Normal distribution and hence derive the estimates of the mean and variance.

7 What is the difference between the method of maximum likelihood and OLS?

Nonparametric Density Estimation for Energy Price Returns

Sitting on the very tip of my chair, feigning interest in the mumbled string of motivational buzz words spouting out of the mouth of an unusually dull director of global risk, it occurred to me that if I looked hard enough, through the gray mist of the incoherent mutterings, there would emerge some shape, some form to their ideas, which as yet my colleagues and I could not perceive. I mused on this thought, toyed with the idea of developing a statistical algorithm that would filter out the noise, revealing the underlying structure. My jocose thoughts were shattered by what was supposed to be the motivational crescendo – we all rose to our feet and clapped our hands somewhat like well-fed seals at feeding time at the local zoo – that is, with not much enthusiasm. Unfortunately, for that individual, there was no form to his ideas, no shape to his plan. Needless to say the listless, MBA-clad, mumbo-jumbo speaking "hot shot" was not head of global risk for very long. However, the experience stuck in my mind and re-stimulated my interest in nonparametric statistical methods, a subset of which, non parametric density estimation, is the subject of this chapter. It introduces nonparametric density estimation as a complementary statistical mechanism for describing energy price returns. It begins by discussing, the simplest nonparametric density estimator – the histogram, how to construct it, and its properties and limitations. This is followed by its generalization into kernel density estimators. Such estimators are often perceived as an improvement over histograms. The chapter discusses why this is so and outlines rules on how to construct them so as to extract all the significant features of price return

data. The chapter ends by giving some tips on how to explain and describe empirical distributions to those not steeped in the language of statistics.

6.1 DESCRIBING ENERGY PRICE DATA WITH HISTOGRAMS

In the previous chapter we discussed fitting parametric probability distributions to energy price return data. This approach assumed we knew something about the underlying (true) probability distribution, and then attempted to fit a good proxy (such as the logistic, normal etc) to it. In one sense by adopting such an approach we are imposing a predetermined form, a shape, onto the data. An alternative approach, which I have found particularly useful, is to encourage the data to tell its story – this can partly be achieved by the use of descriptive statistics; it can also be assisted by nonparametric density estimation discussed in this chapter. The objective of density estimation is to approximate the true probability density of a random variable such as the spot price return of Brent Crude. There are two main types of density estimation: parametric and nonparametric. The distinction between the two is that parametric methods, discussed in Chapter 5, assume prior knowledge about the distribution of the data, while nonparametric methods do not. A histogram is the simplest nonparametric density estimator. It provides information on the nature of the population distribution that generated the sample, the central tendency of the data (median, modes, and so on), the degree of dispersion, and whether or not there are any extreme events or outliers.

6.1.1 Histograms as a probability density estimator

The core idea of a histogram is to locally represent the density implied by sample data by counting the number of observations in a sequence of consecutive sub-intervals, known as bins. Construction of a histogram proceeds by dividing the interval covered by the sample data into equal width bins. For every individual observation that falls into a particular bin a block of size 1 equal to the bin width is placed on top of it. Figure 6.1 shows a typical histogram, in this case the quarterly returns of the WEFA Steam Coal ARA price index over the period March 1996 to December 2003. For explanatory purposes the number of observations in each bin are given at the top of each bar; for example, in the first (far left) bin of Figure 6.1 there is one observation, this value is reported at the top of the corresponding bar.

The histogram is drawn using the actual frequencies rather than relative frequencies. The histogram appears to indicate that the underlying density is unimodal with a moderate degree of kurtosis implied by the relatively
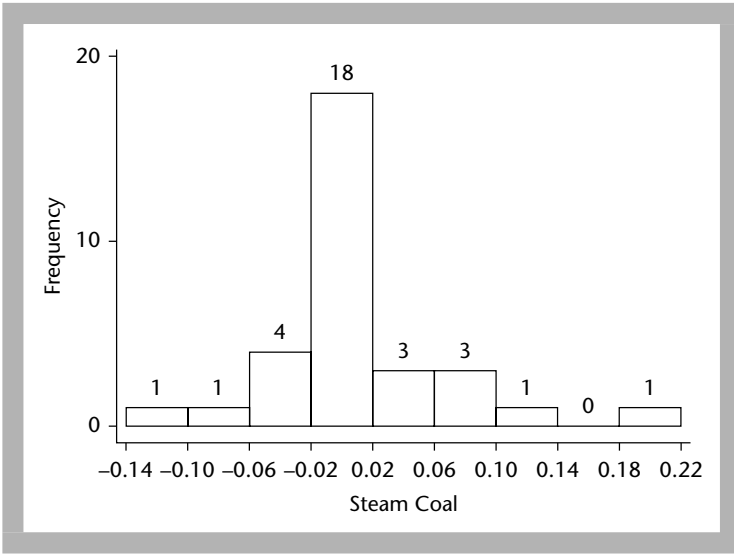
Figure 6.1 Histogram of WEFA Steam Coal ARA quarterly price returns

large number of observations in both tails. In fact the coefficient of kurtosis is 5.7 and the coefficient of skew is around +0.9. In practice if we wish to compare a histogram to a theoretical density we should construct it using relative frequencies. In this case, we would choose a starting point for the first bin say x_0 and a bin width h. Given a sample on $\{x_1, ..., x_N\}$ the histogram estimate, which we denote by $\hat{f}^H(x)$, of some theoretical probability density $f(x)$, is given by

$$\hat{f}^H(x) = \frac{1}{Nh} \quad \text{(number of } x_i \text{ in the same bin as } x\text{)}.$$

Therefore in constructing a histogram we need to consider the bin width h and the starting point x_0.

6.1.2   Choosing the "optimal" bin width

The choice of the bin width has a particularly marked effect on the shape of a histogram. For example, a bin width equal to 0.15 results in the histogram shown in Figure 6.2. This is a completely different estimate of the underlying density than that observed in Figure 6.1. A different picture emerges if we select a bin width of 0.005 as illustrated in Figure 6.3. In this case the histogram appears bimodal and is overly spiky.

The sharp difference between the two histograms raises the question as to what bin width is optimal for providing the best descriptor of the underlying
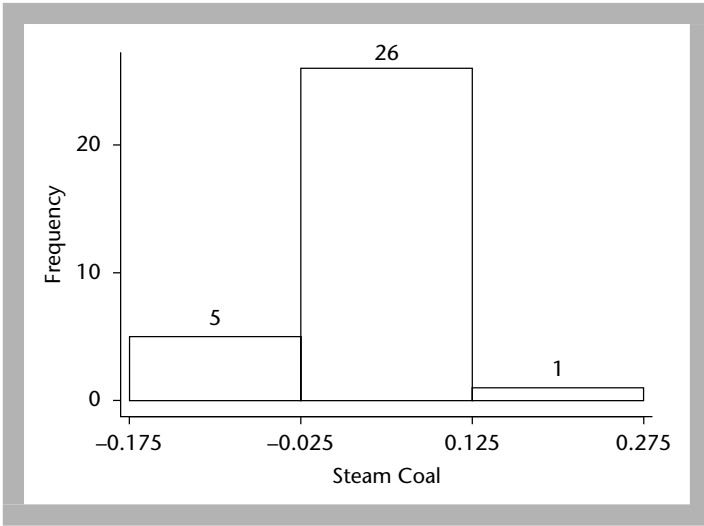
Figure 6.2 Histogram of Steam Coal quarterly price returns
with bin width = 0.15
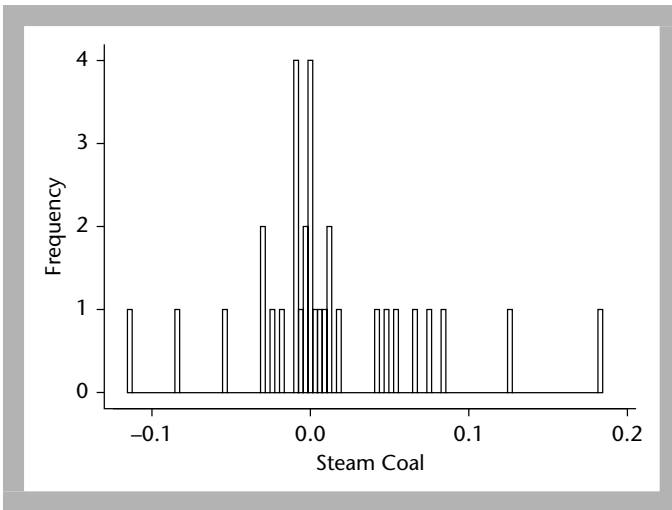
Figure 6.3 Histogram of Steam Coal quarterly price returns
with bin width = 0.005

density? A number of rules of thumb have been proposed to assist with this
issue. Two of the most popular are:

Rule 1: $h = 3.49 \times \sigma \times N^{-(1/3)}$,
Rule 2: $h = 2 \times \text{IQR} \times N^{-(1/3)}$,

where h is the bin width, N the sample size, σ the sample standard deviation
and IQR the inter-quartile range. Figure 6.4 shows two histograms of the
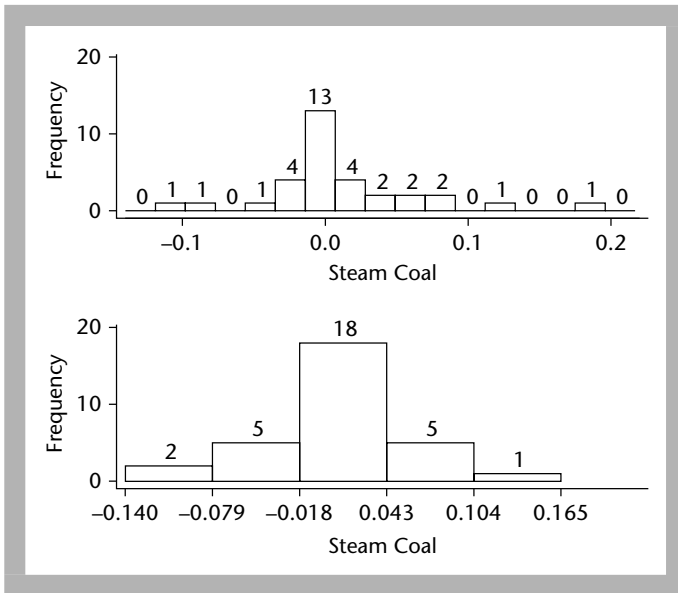
Figure 6.4 Histogram of Steam Coal quarterly price returns
with bin width determined by rule 1 (bottom) and rule 2 (top)

quarterly returns of the WEFA Steam Coal ARA price index. The bottom
histogram calculates the bin width using rule 1. In this case the bin width
is $h = 0.061$. The top histogram uses rule 2 with the bin width value of
$h = 0.021$. Whilst there are some clear differences between the two histo-
grams, the contrast is not nearly as great as that between Figure 6.2 and
Figure 6.3.

Since the size of h determines the number of bins, we could also pre-
specify the number of bins and then work out the required bin width. A
popular way to achieve this is to use the following rule:

Rule 3: Number of bins $= 1 + 3.3 \times \log_{10}(N)$
  This rule works reasonably well provided $N > 15$. Another frequently used
alternative is:
Rule 4: Number of bins $= (2 \times N)^{1/3}$

The bottom histogram, in Figure 6.5, calculates the bin width using rule 3. In
this case there are 6 bins. The top histogram uses rule 4 and contains 4 bins.
Again there are some clear differences between the two histograms.

From the above discussion we see that we should not passively construct a
histogram, without some experimentation with the bin width. This is partic-
ularly important for risk management, as price returns can be highly skewed
and fat tailed – important characteristics that may be missed or exaggerated
by choice of an inappropriate bin width. Too large leads to very big blocks
and thus to a very unstructured histogram. Too small gives a very variable
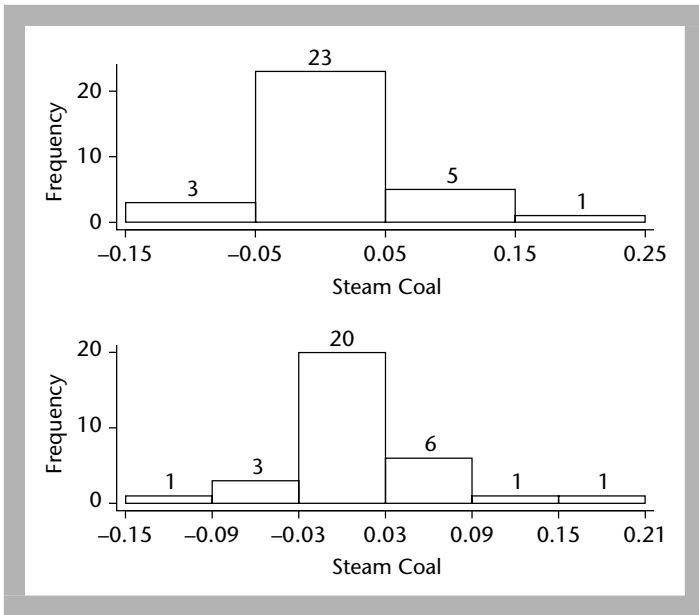
Figure 6.5 Histogram of Steam Coal quarterly price returns
with bin width determined by rule 3 (bottom) and rule 4 (top)

estimate of the underlying density with many unimportant spikes. Unfortunately there is no one best solution for determining bin width. It is always worth trying all four of the rules given here to see whether they are "telling the same story," one which is consistent with the descriptive statistics.

6.1.3   The effect of differing starting points

The choice of the starting point x_0 also has a marked effect on the shape of a histogram. For example, Figure 6.6 shows three histograms for the quarterly returns of the WEFA Steam Coal ARA price index using rule 1. The only difference between the histograms is the value used for x_0. In the top histogram it is set to -0.10, -0.12 in the middle histogram and -0.14 in the bottom histogram. Even though we have used exactly the same data with the same bin width, the histograms give quite different stories of some of the key features of the data: whereas all histograms indicate that the underlying density is unimodal, only the bottom histogram suggests a symmetrical density. Once again, and as was the case with differing values of h, we find ourselves with three completely different estimates of the underlying density.

The property of histograms for representing the same data quite differently strongly contradicts with the goal of presenting pertinent features of the data. It also conflicts with the goal of nonparametric statistics to "let the
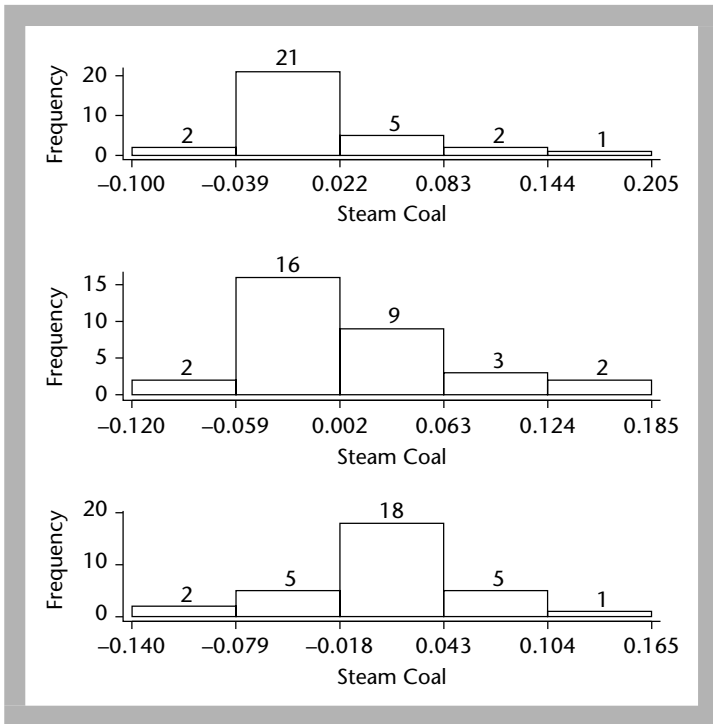
Figure 6.6 Histogram of Steam Coal quarterly price returns
with three different starting points

data speak for themselves." We have seen that histograms of the same data with the same bin width need not be identical and that the shape depends heavily on both the starting point and the bin width. Too large a bin width results in a unstructured histogram, too small a bin width results in an unstable noisy histogram. Obviously, the same data speak quite differently out of the different histograms. Despite these limitations it is clear that histograms should be viewed by the risk manager as more than a simple and convenient tool for the graphical representation of empirical data. They are a reliable tool for estimating the underlying probability distribution for price returns.

6.1.4   The average shifted histogram

To overcome some of the problems with basic histograms, the concept of an average shifted histogram (ASH) has been developed. This approach computes a number of histograms using the same bin width but different starting points and then takes the average over the histograms to characterize the
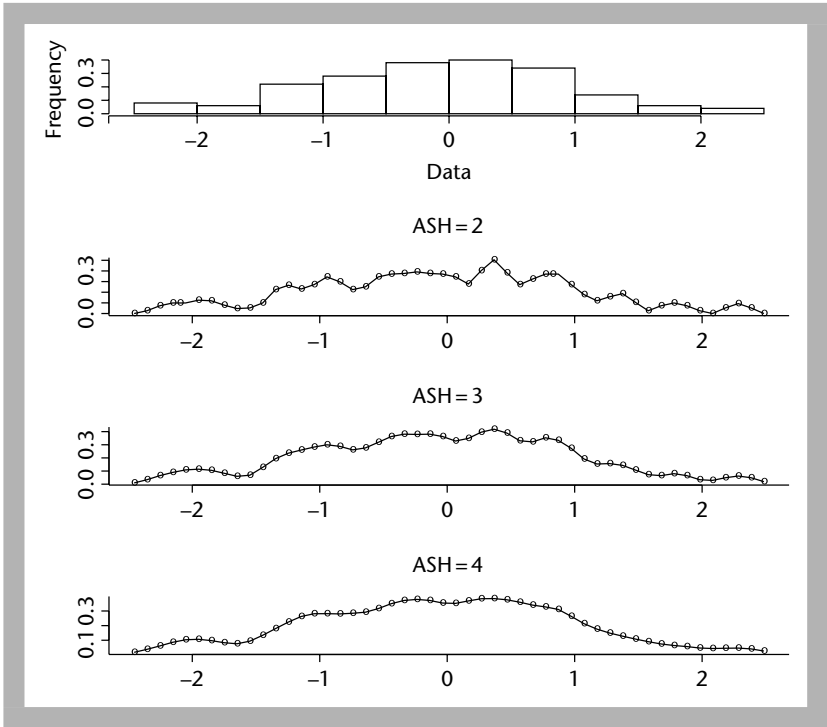
Figure 6.7 Average shifted histograms for 100 observations from a normally distributed random variable

data. This removes the dependence of the histograms' shape on the starting points. Figure 6.7 (top) shows the histogram for 100 observations from a standard normal random variable. Whilst in this case the histogram appears fairly symmetric, as a density estimate, it is not very smooth. Underneath (second from top) is an ASH constructed using two histograms, followed by an ASH using 3, and then 4 histograms. Notice how the ASH histograms appear increasingly smoother as the number of histograms used increases. All three ASH diagrams capture the characteristics of the underlying probability distribution including symmetry very well. Figure 6.8 shows ASH diagrams for the quarterly returns of the WEFA Steam Coal ARA price index. The resulting diagram is freed from the problems associated with dependence on differing starting points illustrated in Figure 6.6. All three diagrams in Figure 6.8 indicate a fairly symmetrical distribution with an average return slightly greater than zero.

Whilst the ASH is often viewed as a histogram with a smaller bin width, you should note that this is not quite the case because, as we have seen, the shape of ordinary histograms is dependent on the starting points, ASHs are not. In addition, the discontinuities of a histogram estimate are not due to the underlying density, they are only an artifact of the chosen bin widths.
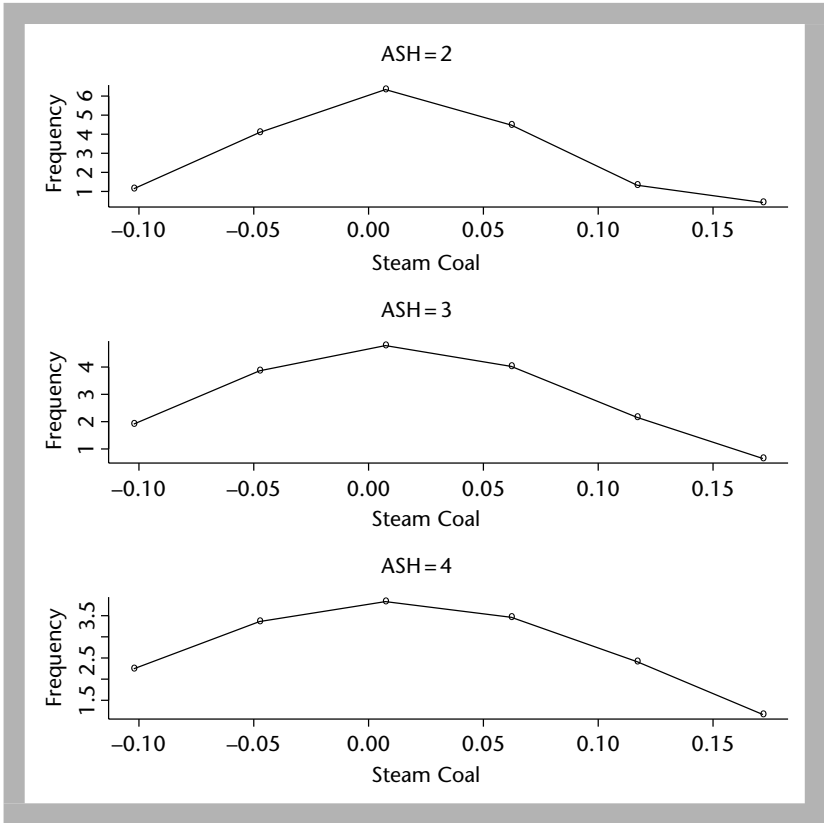
Figure 6.8 ASH histograms for Steam Coal quarterly price returns

These discontinuities can make it difficult, without experience, to grasp any underlying structure.

The concept of smoothing using averages is a central idea in statistics. By using varying starting points ASH may better be able to extract structural elements of complexity from the patterns of random variation apparently inherent in energy price returns. ASHs are thus designed to simultaneously estimate and model the underlying structure. They are actually kernel density estimators, a subject we turn to in the next section.

6.2 KERNEL DENSITY ESTIMATION

A kernel density estimator is defined as

$$\hat{f}_h^{\text{K}}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right),$$

where N are the number of observations in the sample and h is known as the bandwidth (similar to the bin width in a histogram). The estimator calculates the density at the center of the weighted xs found within a rolling interval or "window." In other words a kernel function is generated around each data point x_i in the sample. An estimate of the underlying density function is obtained by adding these kernel functions together and scaling the sum to equal 1. Kernel density estimators differ from histograms both in allowing the windows to overlap and by the possibility of using different weighting schemes on the xs. The function $K(.)$, which determines the weights, is termed the "kernel" and is usually chosen as a symmetric probability density satisfying the condition

$$\int_{-\infty}^{\infty} K(u)\,du = 1.$$

6.2.1   Choosing the appropriate bandwidth

In order to illustrate the calculation of a kernel density consider the uniform kernel for which

$$\hat{f}_h^K(x) = \frac{1}{2Nh}(\text{number of sample observations } x_i \text{ that fall in } [x-h, x+h]).$$

Figure 6.9 shows a density plot of coal steam using this uniform density and with $h = 0.025$. The kernel density estimate appears smoother than the histogram estimates discussed in the previous section. However it does suggest, at first glance, that the data is at least bimodal. This may be an artifice of the data. We can attempt to eliminate these by increasing h.

Figure 6.10 shows the result of setting h equal to 0.5 (top), 0.2 (middle), and 0.01(bottom). It is evident that the larger h the smoother the resulting density plot. Conversely, the smaller h, the coarser and more jagged the density plot. It is therefore critically important to choose an appropriate value for h. Too large a value results in an over-smooth estimate and too small results in an estimate that contains too much noise.

The amount of structure observed is therefore critically dependent on the bandwidth. A number of rules of thumb have been proposed to assist with this issue. Three of the most popular are:

Kernel Rule 1: $h = \sigma \times N^{-1/5}$
Kernel Rule 2: $h = 0.79 \times \text{IQR} \times N^{-1/5}$
Kernel Rule 3: $h = 1.06 \times \min\left\{\sigma, \dfrac{\text{IQR}}{1.34}\right\} \times N^{-1/5}$
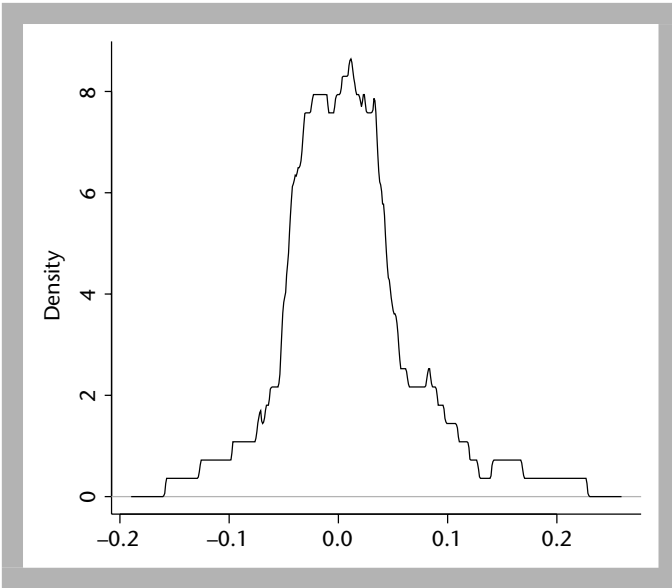
Figure 6.9 Rectangular density kernel plot of Steam Coal with $h = 0.025$ for Steam Coal quarterly price returns
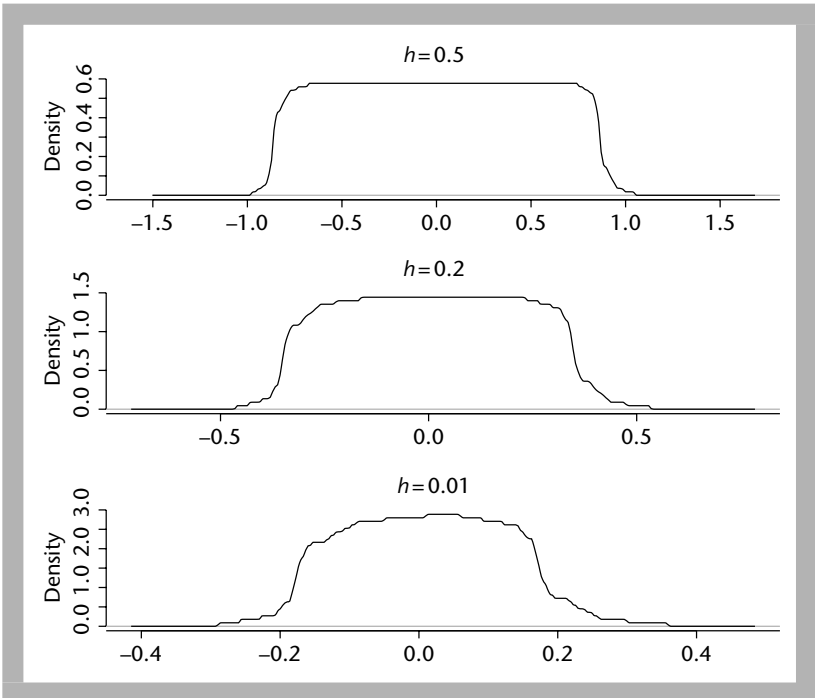
Figure 6.10 Rectangular density kernel plot of Steam Coal for various bandwidths
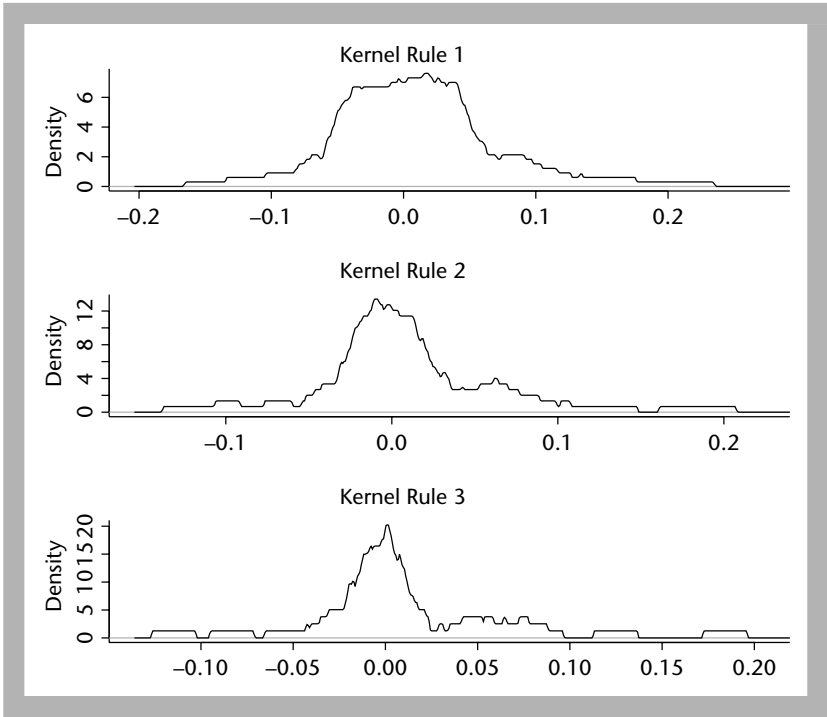
Figure 6.11 Rectangular density plot of Steam Coal for using kernel rules 1, 2, and 3

Kernel rule 1 and kernel rule 3 work best when the underlying distribution is close to the normal distribution. Figure 6.11 shows three kernel density plots of the quarterly returns of the WEFA Steam Coal ARA price return using the above rules with a uniform kernel. The top diagram uses rule 1 with $h = 0.0296$, the middle diagram rule 2 with $h = 0.0134$, and the bottom diagram rule 3 with $h = 0.007$.

Unfortunately, as with histogram bin width, there is no one best solution for determining h. It is always worth trying all three of the above rules alongside experimentation and then make an assessment as to which is most consistent with the descriptive statistics and any prior knowledge.

6.2.2   Choosing an appropriate kernel

Construction of a kernel density requires the selection of a suitable bandwidth and a choice of kernel. Which of the numerous probability distributions or other functions should we choose? Fortunately, the choice of kernel is not nearly as important as the choice of bandwidth, this is illustrated in Figures 6.12 and 6.13, which show density plots of WEFA Steam Coal ARA
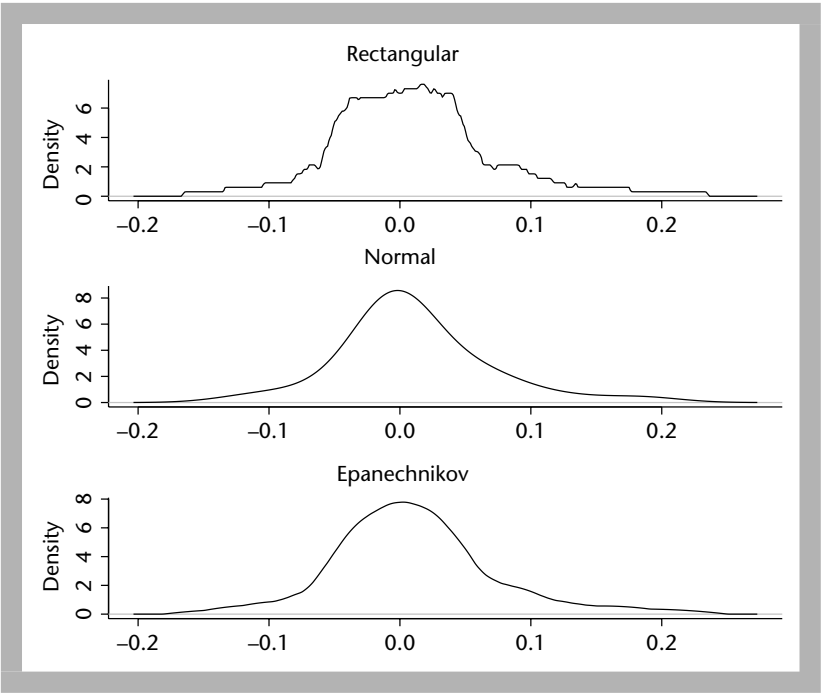
Figure 6.12 Density plot of Steam Coal using rules 1 and Rectangular, Normal, and Epanechnikov kernels
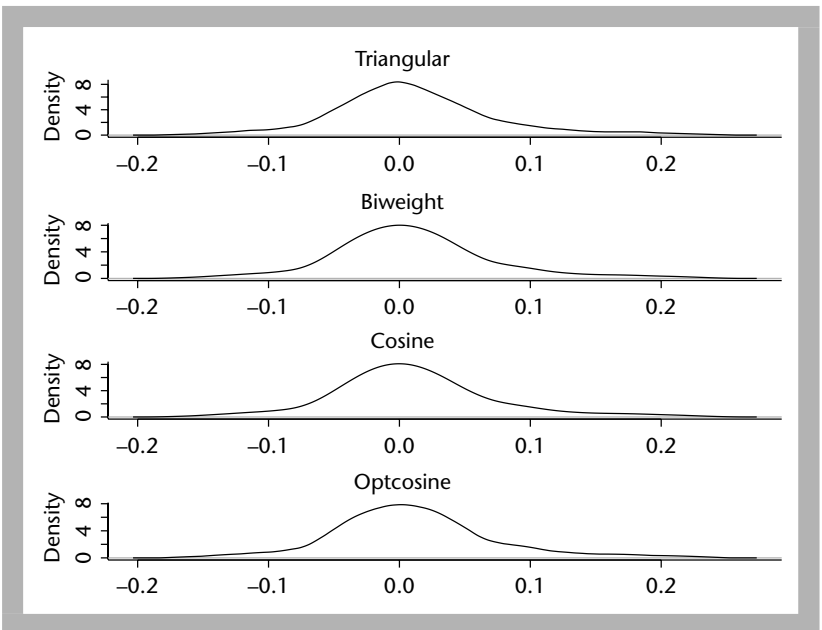
Figure 6.13 Density plot of Steam Coal using rules 1 and Triangular, Biweight, Cosine, and Optcosine kernels

price return for different kernels using rule 1. With the exception of the uniform/rectangular kernel (top diagram in Figure 6.12), there is little to differentiate between the others.

Whilst the choice of kernel is not nearly as important as the choice of bandwidth we should not passively construct a kernel density estimate without some experimentation. It is always worth trying various kernels to see whether they are "telling the same story," one which you feel is consistent with the given data. Alas, there is no automatic method for determining the optimal kernel.

6.3   EXPLAINING EMPIRICAL DISTRIBUTIONS TO NONSTATISTICAL PEOPLE

There is something about the language of statistics that is somehow repellent to nonstatisticians. Maybe it has to do with distasteful experiences in Statistics 101, maybe not. Either way, the quickest way to lose the attention
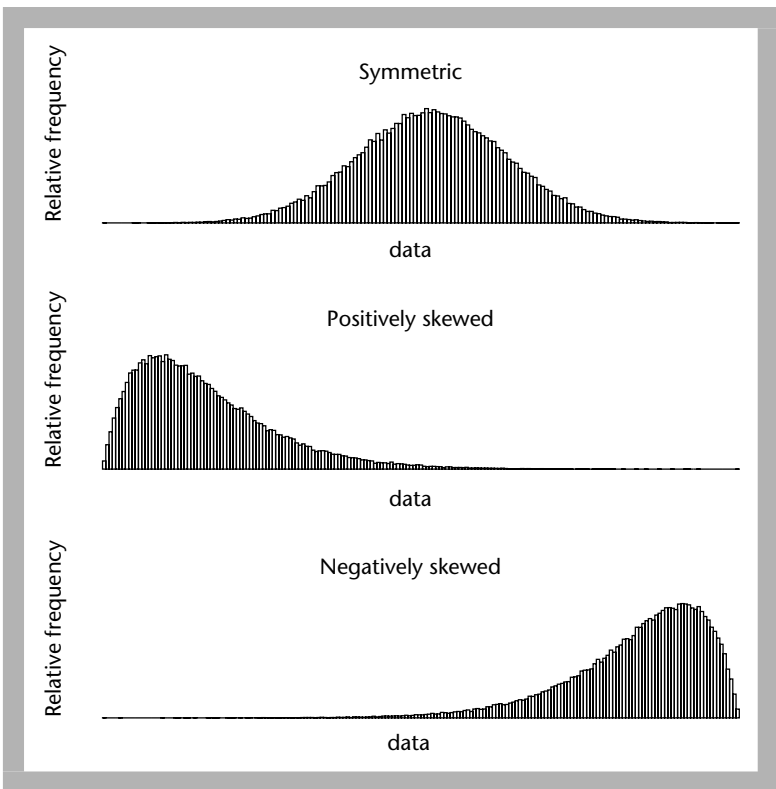
Figure 6.14 Useful diagrams for discussing symmetry with nonstatistical individuals
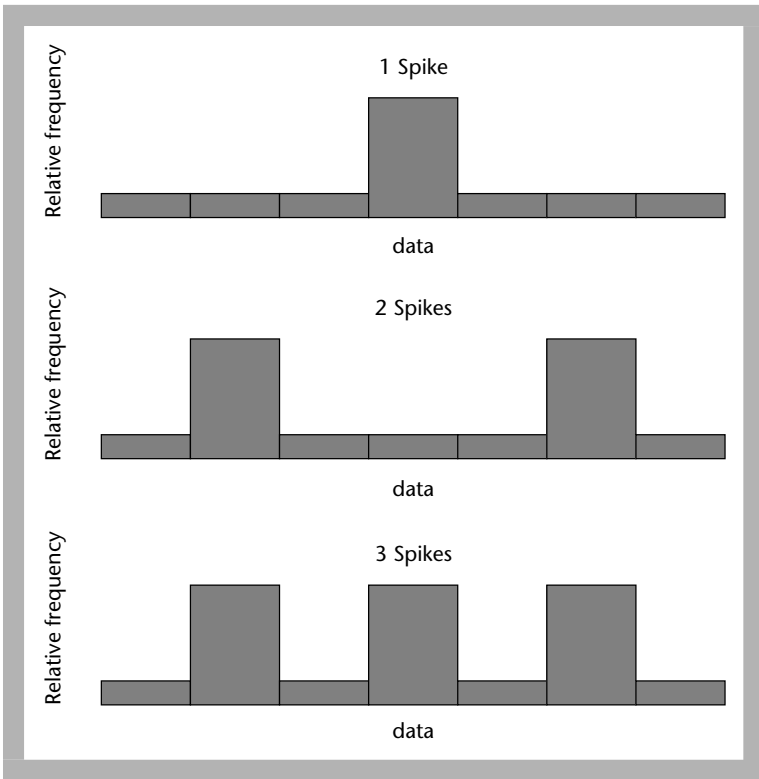
Figure 6.15 Useful diagrams for discussing peaks in distributions with nonstatistical individuals

of most senior executives is to talk in the language of statistics. It is one of our responsibilities as risk managers, risk analysts, and quantitative analysts to pass on essential quantitative information, free from the erudite haze than has traditionally surrounded mathematical disciplines. Whilst distribution names, such as Normal, Logistic, and Laplace are useful to ensure compact communication between ourselves, they are useless for conveying information to those not steeped in the discipline. This is because their names do not carry with them any inherent information about the characteristics they attempt to capture. The distribution name tells the nonstatistical person nothing about the shape.

A long time ago, in a causal conversation with a Chief Investment Officer, I happened to mention that the returns for a particular product were highly non-normal. The individual took non-normal to mean abnormal and became extremely agitated at the possibility of such returns occurring "on their shift." It would have been far better for me to have mentioned that the returns were positively skewed with a large probability of a positive payoff! A simple diagram of the situation may have been sufficient. Effective risk
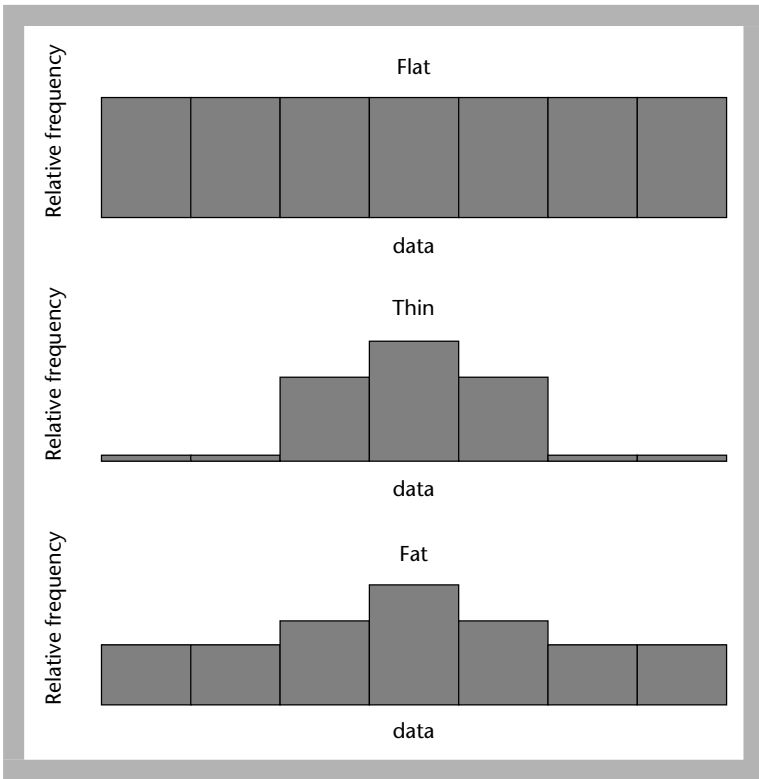
Figure 6.16 Useful diagrams for discussing uniform, thin- and fat-tailed distributions with nonstatistical individuals

management is not simply about quantitative methods. Success requires communication to senior managers and co-workers in ways which insulate and isolate them from the rigors and details of statistical and mathematical methods and language. In talking to nonstatisticians I always remember the astute words of the mythical statistician Cory Lation:[1]

Whoever invented statistical terms
Had a head that was stuffed with worms.
All these new words are so much junk,
And if I don't learn them, I'm really sunk.

In describing probability distributions to senior management and other non-statistical people, I usually refer to shape characteristics (skew, fat tails, symmetric) rather than names of distributions or worse parameters of the distributions. For example when describing a symmetric distribution (say, the Normal distribution) it is preferable to refer to the mean and standard deviation than the parameter names of mu (μ) and sigma (σ); but it is even

better to refer to shape characteristics. In fact, the old saying that a picture is worth a thousand words is particularly relevant in risk management. In terms of symmetry the simple diagrams of Figure 6.14, are usually sufficient to convey to nonstatistical people what this concept implies in terms of price returns or other risk factors. Figure 6.15 is another diagram that I find useful when discussing modes or spikes. Finally, Figure 6.16 is useful when discussing uniform (rectangular) distributions which I often term flat distributions, and concepts such as fat and thin tailed. Always remember that many professional people are not versed in the language of statistics, nor do they particularly wish to be.

6.4   FURTHER READING

Silverman (1986), Marron and Nolan (1988), and Scott (1992) provide very accessible introductions to nonparametric density estimation. Detailed discussion of bandwidth selection are given in Marron (1989), Park and Turlach (1992), and Turlach (1993). Smoothing techniques particularly relevant to energy risk modeling are given in Härdle and Scott (1992), and Wand and Jones (1995). In addition Härdle (1991) discusses implementations in the statistical language *S*, which can with minimal changes be ported into R or a spreadsheet package.

Härdle, W. (1991) *Smoothing Techniques, With Implementations in S*, Springer, New York.

Härdle, W. and Scott, D. (1992) Smoothing in by weighted averaging using rounded points, *Computational Statistics*, 7, 97–128.

Marron, J. (1989) "Comments on a data based bandwidth selector," *Computational Statistics & Data Analysis*, 8, 155–70.

Marron, J. S. and Nolan, D. (1988) "Canonical kernels for density estimation," *Statistics & Probability Letters*, 7(3), 195–9.

Park, B. U. and Turlach, B. A. (1992) Practical performance of several data driven bandwidth selectors, *Computational Statistics*, 7, 251–70.

Scott, D. W. (1992) *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, New York and Chichester.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*, Vol. 26 of Monographs on Statistics and Applied Probability, Chapman and Hall, London.

Turlach, B. A. (1993) *Bandwidth Selection in Kernel Density Estimation: A Review*, Discussion Paper 9307, Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin.

Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*, Vol. 60 of Monographs on Statistics and Applied Probability, Chapman and Hall, London.

6.5   REVIEW QUESTIONS

1 What are the strengths and weakness of histograms as density estimators?

2 Explain the impact of bin width on the density estimate of a histogram.

3 Why can changing the starting point of a histogram have a drastic impact on its shape?

4 What are the advantages of ASH over ordinary histograms?

5 Explain why ASH is a kernel estimator.

6 Describe the procedure for fitting a kernel density to price returns. What are the advantages and pitfalls of this approach?

7 Discuss the pitfalls of using statistical language when discussing distributional shapes with nonstatistical individuals. What can you do to improve the situation?

Correlation Analysis

To this point, we have dealt almost exclusively with problems of estimation and statistical inference about a parameter of a probability distribution or characteristic of a sample. Another important element of applied statistical modeling of energy risks concerns the relationship between two or more price or other variables. Generally, a risk manager will be interested in whether above (below) average values of one variable tend to be associated with above (below) average values of the other variable. Take for example, a risk manger working for a petroleum refinery, who for hedging purposes, is interested in knowing the relationship between the spot price of Brent Crude and the future price of diesel fuel. If the risk manager simply assumes crude oil and diesel fuel prices always move in tandem, the company will be exposed to the price risk if this relationship breaks down. If on the other hand, the closeness of the two indices is defined in terms of a correlation coefficient, then the manager at least has some rudimentary way of assessing whether or not the relationship exists and its strength.

This chapter is devoted to the elementary techniques of correlation analysis – the study of relationships among a number of random variables such as price and returns. It begins with definitions and properties of correlation, continues with various ways to calculate correlation and ends with a discussion of correlation and causation.

7.1 UNDERSTANDING CORRELATION

A correlation coefficient measures the extent to which two variables are related to each other. Of course in principle, there are any number of ways in which two variables might be related, therefore it is useful to postulate some functional form. A correlation coefficient assumes a linear relationship between the two variables. It takes values between -1 and $+1$. When the
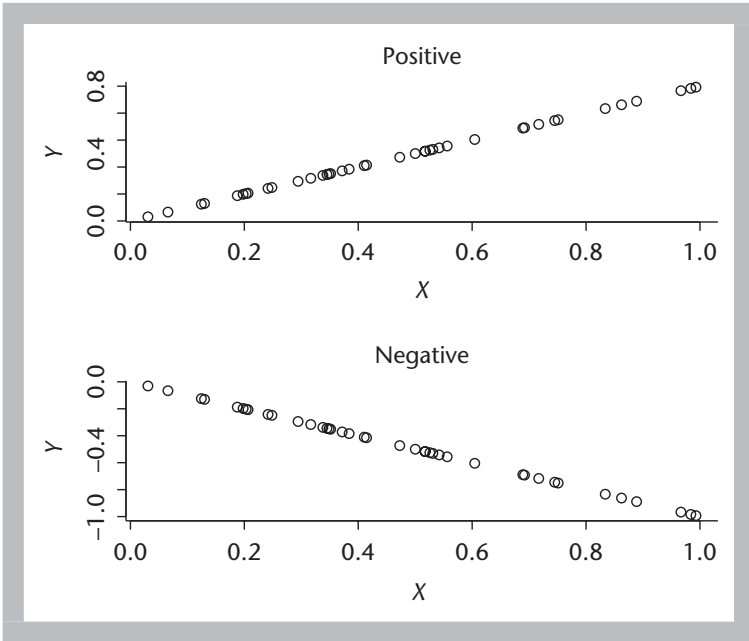
Figure 7.1 Scatter plot of two variables with perfect positive
(top) and negative (bottom) correlation

correlation coefficient is equal to +1 (−1) there is perfect positive (negative)
correlation, and when it is equal to zero there is no correlation. Figure 7.1
shows a scatter plot for the situation of perfect positive (top) and negative
correlation (bottom). Notice how in this case the points lie along a straight
line, which is upward sloping in the case of positive correlation and down-
ward sloping in the case of negative correlation. Figure 7.2, shows the case
where the correlation is equal to +0.70 (top diagram) and −0.70 (bottom
diagram). Whilst the correlation is not perfect we can clearly see that for
a positive correlation (top diagram) high values of one variable are associ-
ated with high values of the other variable, and low values of each variable
are associated with one another. For a negative correlation (bottom dia-
gram) high values of one variable are associated with low values of the
other variable. A correlation of zero indicates no linear relationship between
the variables. As shown in Figure 7.3, zero correlation leads to a random
looking series of points with no discernable pattern.

Correlation coefficients provide a useful scale against which the closeness
of the relationship between two variables can be measured. Therefore, the
closer is the coefficient to its limit of ±1, the stronger the linear relation-
ship between the two variables. Given two random variables X and Y with
standard deviation σ_X and σ_Y respectively; The correlation coefficient can
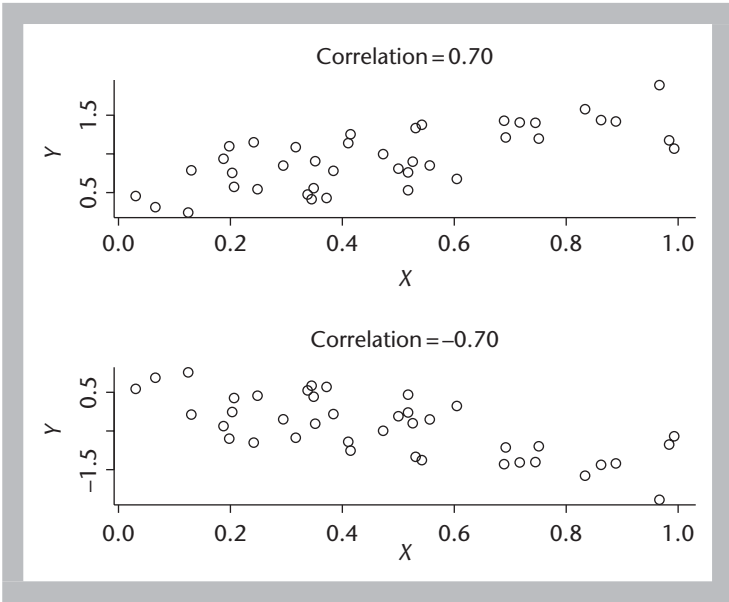
Figure 7.2 Scatter plot of two variables with positive correlation of 0.70 (top) and negative correlation of −0.70 (bottom)
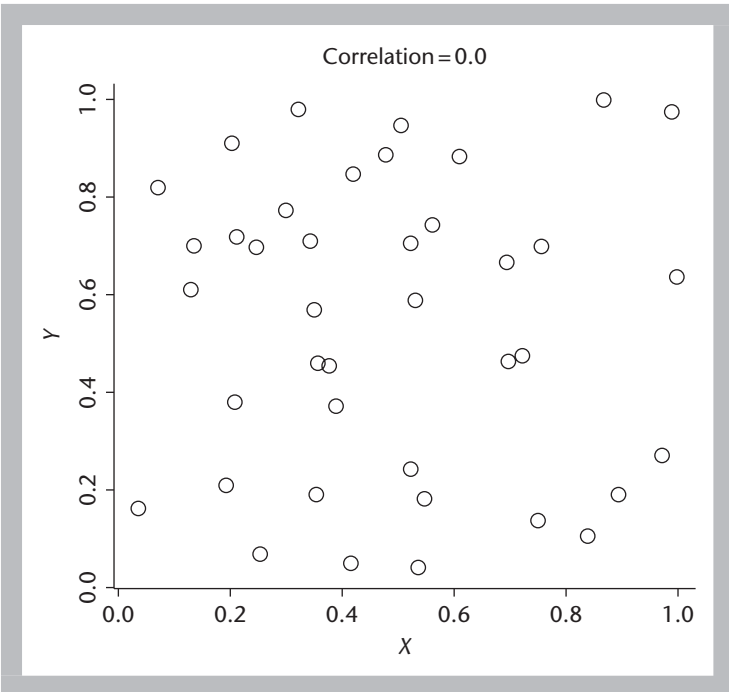
Figure 7.3 Scatter plot of two variables with zero correlation

be calculated as

$$\rho = \frac{\text{Covariance } (X, Y)}{\sigma_X \sigma_Y}$$

7.2   CORRELATION AND HEDGING

Frequently one hears the claim that to hedge a unit of a spot price position in an energy product one should take on an opposite position of 1 unit in the corresponding futures contract. This strategy assumes the spot and future markets move in tandem. If this is not the case the hedging strategy will be subject to basis risk. An alternative approach to determining the hedge is to note that given the spot price S and futures price F the variance of a portfolio of h units of F and one unit of S can be calculated as

$$\sigma^2 = \sigma_S^2 + h^2 \sigma_F^2 - 2 \times h \times \text{Covariance}(F, S).$$

The question now is what value should h take? If we assume S and F always move in tandem then it is arguable that it should have a value of 1. However, since we know this is unlikely to be the case and because variation in this portfolio is associated with price risk we could choose h to minimize σ^2. If we choose this route then we see that

$$\frac{\partial \sigma^2}{\partial h} = 2h\sigma_F^2 - 2 \times \text{Covariance}(F, S) \quad \Rightarrow \quad h = \frac{\text{Covariance}(F, S)}{\sigma_F^2}.$$

Notice that if

$$\sigma_F^2 = \sigma_S^2 \quad \text{then } h = \frac{\text{Covariance}(F, S)}{\sigma_S \sigma_F} = \rho,$$

which is of course the correlation between S and F.

Example 7.1   Optimal hedge ratio for Jet fuel

The basic principles of hedging can be used for many commodities for which no futures contract exists, because often these are similar to commodities that are traded. Jet fuel prices exhibit substantial volatility with periods of sustained rising (falling) prices. Rising prices place pressure on airline costs hindering their ability to maintain positive cash flows. Whilst Jet fuel is hedgable, there is not a perfect hedge available and alternatives such as crude oil or heating oil futures/forwards must be used. Figure 7.4 presents a time series plot of the daily Jet fuel spot price and 3-month Brent Crude
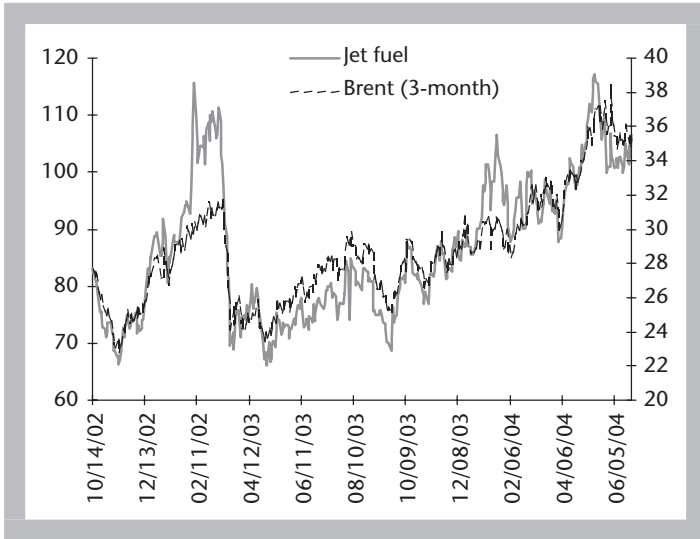
Figure 7.4 Time series plot of Jet fuel spot and 3-month Brent Crude future prices

future. As we might expect the two prices appear to be closely related, with a correlation of 0.9; The correlation between log returns is slightly lower at 0.712. Thus h, the number of contracts required for hedging, can be calculated as

$$h = \frac{\text{Covariance(Brent 3-month, Jet fuel spot)}}{\text{Variance(Jet fuel spot)}} = 1.042.$$

Clearly interrelationships between futures, forwards, and spot prices can have a significant impact on business performance and the overall risk profile; Therefore recognition of the correlation between prices is necessary for energy players to manage risk better. In practice there are several ways to calculate a correlation coefficient. We discuss some of these in the following sections.

7.3   PEARSON PRODUCT MOMENT CORRELATION COEFFICIENT

Given a sample of N paired observations each on two continuous random variables $X = \{x_1, x_2, ..., x_N\}$ and $Y = \{y_1, y_2, ..., y_N\}$, the product moment correlation coefficient can be calculated as

$$\rho = \frac{1/(N-1) \sum_{i=1}^{N} (x_i - \bar{X})(y_i - \bar{Y})}{S_X S_Y},$$

where S_X and S_Y are the sample standard deviations and $\bar{X}$ and $\bar{Y}$ are the sample means. If we let $\tilde{x}_i = (x_i - \bar{X})$ and $\tilde{y}_i = (y_i - \bar{Y})$ we can rewrite the formula slightly modified as

$$\rho = \frac{1/(N-1)\sum_{i=1}^{N}\tilde{x}_i\tilde{y}_i}{S_X S_Y}.$$

Since the $\tilde{x}$'s and $\tilde{y}$'s are the deviations from the mean of X and mean of Y respectively it is immediately apparent that if large positive values of X are associated with negative values of Y then the product moment correlation coefficient must be negative. On the other hand if positive values of X are associated with positive of Y the coefficient is positive. A simple function written in R to calculate the Pearson correlation coefficient is given below

```
pm.corr<-function(x,y)
{
xx=x-mean(x)
yy=y-mean(y)
sdx=sd(x)
sdy=sd(y)
N=length(x)
(1/(N-1)*sum(xx*yy))/(sdx*sdy)
}
```

In fact this correlation coefficient is available in most spreadsheet and statistical packages. For example, it can be accessed in Excel using the function `Correl(data_range_1, data_range_2)`.

Example 7.2   Product moment correlation between unemployment and Gross National Product (GNP)

From Chapter 1 we know that the general economic environment can have an impact on the risks facing energy companies. Consider the situation where higher than expected energy prices lead to a slowing of GNP growth and productivity, which in turn lead to slower wage growth and an increase in the unemployment rate. As part of the ongoing background analysis of potential risks to the energy business a risk manger may be interested in the relationship between GNP and unemployment. Let us pursue this idea briefly using the R statistical package. The R data frame `longley` contains seven macroeconomic variables observed yearly from 1947 to 1962. A scatter plot of the two variables GNP and unemployment is shown in Figure 7.5. A casual glance at the scatter diagram appears to indicate a positive relationship between these two variables. High values of one variable
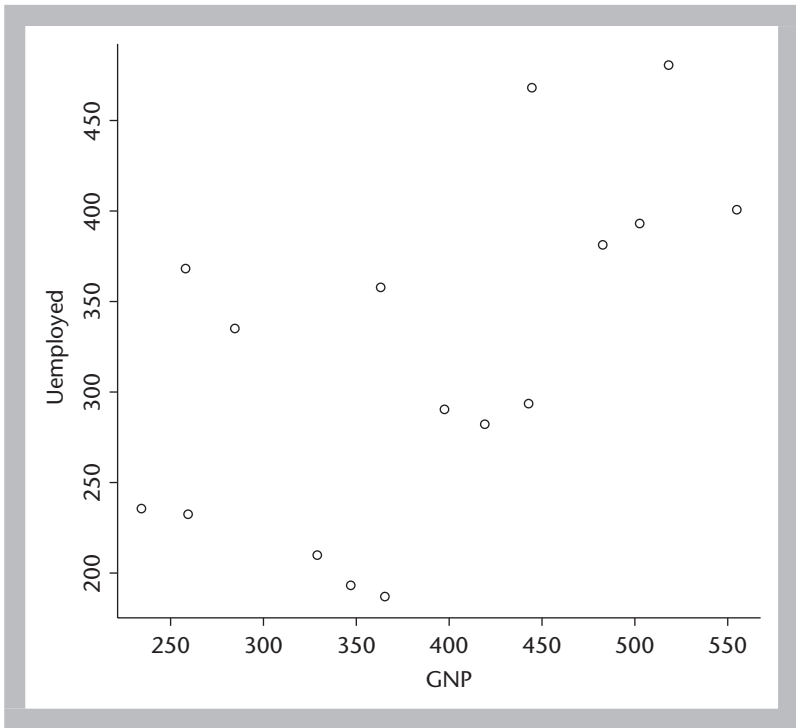
Figure 7.5 Scatter plot of Gross National Product (GNP) and unemployment from data frame Longley

are associated with high values of the other variable. What is the product moment correlation between the level of GNP and unemployment? We can use the previously defined function `pm.corr()` to find out.

```
> pm.corr(longley[,"GNP"],longley[,"Unemployed"])
[1] 0.604261
```

The correlation coefficient is approximately 0.60.[1] How should we interpret this value? First it is positive, and hence, appears to indicate a positive association between the two variables. Second as a general rule of thumb an absolute correlation in the range of 1.0 to 0.7 indicates that the two variables are strongly correlated, 0.7 to 0.3 as moderately correlated and less than 0.3 as indicative of very little correlation. So at first glance with a product moment correlation coefficient of 0.60 it appears that the two variables are moderately and positively associated.

One of the key assumptions underlying the product moment correlation coefficient is that the two variables used in its calculation are jointly normally distributed. For this reason it is known as a parametric correlation

coefficient. Parametric correlation coefficients make an assumption about the joint distribution of the correlated variables. If we do not wish to make this assumption an alternative measure of correlation such as the Spearman rank correlation coefficient can be used.

7.4   SPEARMAN RANK CORRELATION COEFFICIENT

The product moment correlation coefficient is sensitive to extreme observations. Moreover, tests based on it rely for their validity on the assumption that the variables are from the bivariate normal distribution. The Spearman rank correlation coefficient is more robust to extreme values and is also nonparametric because it does not make any assumption about the joint distribution of the variables. It calculates the correlation between two variables using the ranks of the original data. Given a sample of N paired observations each on two continuous random variables $X = \{x_1, x_2, ..., x_N\}$ and $Y = \{y_1, y_2, ..., y_N\}$, provided there are no tied ranks, the Spearman rank correlation coefficient can be calculated as: $\rho = 1 - [6 \sum_{i=1}^{N} d_i^2 / (N^2 - 1)N]$, where d_i are the differences of the ranked pairs of X and Y. The following function written in R can be used to calculate the rank correlation between two variables

```
rank.corr<-function(x,y)
{
rank.x=length(x)+1-rank(x)
rank.y=length(y)+1-rank(y)
d=sum((rank.x-rank.y)^2)
N=length(rank.x)
1-((6*d)/((N^2-1)*N))
}
```

Example 7.3   Rank correlation between unemployment and GNP

Returning to Example 7.2 we see that

```
> rank.corr(longley[,"GNP"],longley[,"Unemployed"])
[1] 0.6382353
```

This value of 0.64 is reassuringly close to the estimate of 0.60 given by the product moment correlation coefficient.

7.5   SPURIOUS CORRELATION

Up to this point we have been content to take the estimate of correlation between two variables as indicative of an actual relationship. However, in interpreting the correlation coefficient in this way we need to exercise extreme caution so as to avoid coming to the wrong conclusion based on spurious correlation. As an example of a spurious correlation consider European storks which breed over parts of central Europe. In such areas there is an increase in the number of new babies born in spring, precisely when the storks appear and begin nesting. If we were to calculate the correlation between babies born and the appearance of the storks it would be high but spurious.

***Example 7.4   Spurious correlation between the level
unemployment and level of GNP***

In Example 7.2 we found a correlation of around 0.6 between the level of GNP and unemployment. Thus, at a first look it appears as if there is a weak positive correlation between the level of GNP and unemployment. Does this make sense? The implication is that as the level of GNP rises so does the level of unemployment. Surely economic growth should have the opposite effect? As the economy grows richer the unemployment level should in some sense fall. The problem here is two fold. First we should really look at real GNP (that is GNP adjusted for inflation – `GNP.deflator`) and second, because both variables grow over time, we should use their rates of change rather than their levels. Using the log rates of change instead of the actual levels we have

```
> logChange.GNPdeflator=c(0.064161944,-
0.003395589,0.014631662,0.072190732,0.019558009,0.009132484,
0.010050336,0.011928571,0.033044795,0.035684537,0.021898685,
0.016114941,0.014109582,0.013049337,0.010318234)
> logChange.Unemp=c(-0.013245227,0.459736044,-0.094197274,-
0.467797768,-0.082905305,-0.032617305,0.648865554,-0.208714888,-
0.028643244,0.039602257,0.466463652,-
0.205095481,0.030477566,0.200971291,-0.181822306)
```

Figure 7.6 shows time series plots and a scatter plot of the variables. The scatter plot (bottom) has a very slight downward slope, indicative of negative correlation. The product moment correlation is

```
> pm.corr(logChange.GNPdeflator,logChange.Unemp)
[1] -0.4534178
```
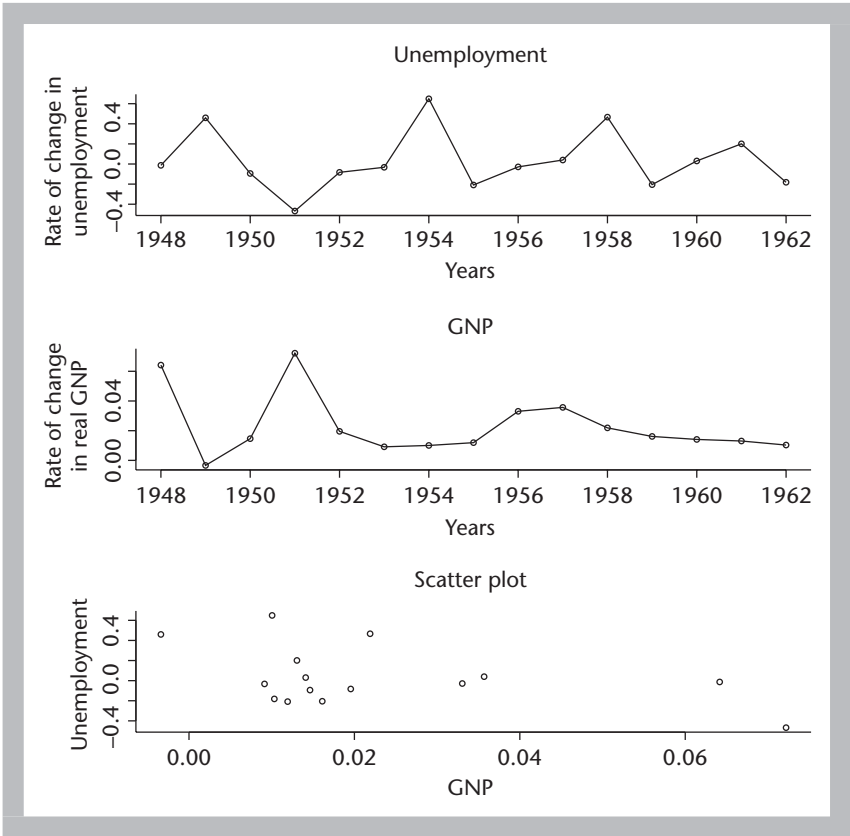
Figure 7.6 Time series plots and a scatter plot of the rate of change in real GNP and the rate of change in unemployment

and the rank correlation is

```
> rank.corr(logChange.GNPdeflator,logChange.Unemp)
[1] -0.2107143.
```

Both values are negative, with the rank correlation coefficient almost half that of the product moment coefficient. What can we conclude? If anything, the relationship between the rate of change in GNP and the rate of change in unemployment is negative – Growth in real GNP is associated with negative rates of change in unemployment.

The key idea derived from this discussion is that measuring and using correlation can be fraught with difficulty. One has to be careful that the data are in the correct format at the start of the analysis. Table 7.1 shows the correlation between the log returns for different types of New York gasoline, Unleaded premium (Non-Oxy) denoted by GUPNO, Unleaded

Table 7.1 Correlation table of log returns of four types of gasoline

	GUPNO	GURNO	GPO	GRO
GUPNO	1.0000			
GURNO	0.9483	1.0000		
GPO	0.9579	0.9114	1.0000	
GRO	0.9537	0.9537	0.8810	1.0000

regular (Non-Oxy) denoted by GURNO, Unleaded premium (Oxy) denoted by GPO and Unleaded regular (Oxy) denoted by GRO. As one might expect, the correlations are all very high and close to 1.

7.6   THE KENDALL TAU COEFFICIENT

The Kendall Tau coefficient is a measure of the strength of the relationship between two variables. It is in this sense similar to a correlation coefficient. Given paired observations on X and Y, each variable is placed in rank order with a value of 1 for the lowest, 2 for the next observation and so on. Like other correlation coefficients it takes values between -1 and $+1$. Tau can be estimated in R using `cor.test()`. However in practice there is little difference between the value obtained using Tau and the estimate of a Spearman rank correlation coefficient. For this reason, it is little used in risk management. Nevertheless, it is always worth calculating, not least for comparison with the value obtained for the Spearman rank and Pearson coefficients.

7.7   CONFIDENCE INTERVALS FOR THE CORRELATION COEFFICIENT

The computation of a confidence interval for a correlation estimate is complicated by the fact that its sampling distribution is not normally distributed. One easy-to-implement solution is to use the Fisher z transformation process. This process involves the following steps:

1. Transform the correlation estimate $(\hat{\rho})$ into a Fisher z score $(\hat{z})$ using the formula

$$\hat{z} = 0.5 \ln\left(\frac{1 + \hat{\rho}}{1 - \hat{\rho}}\right)$$

2. Compute a confidence interval for $\hat{z}$

3. Convert the confidence interval for $\hat{z}$ into a confidence interval for $\hat{\rho}$

Let's use Example 7.1 to illustrate the idea. As we expected, there is a positive relationship between these two variables, the larger the daily return on Brent Crude the higher the return on Jet crude. The correlation based on $N = 443$ observations is 0.712. The problem is to compute a 95% confidence interval given that $\hat{\rho} = 0.712$. As a first step we calculate the Fisher z score as

$$\hat{z} = 0.5 \ln\left(\frac{1 + \hat{\rho}}{1 - \hat{\rho}}\right) = 0.5 \ln\left(\frac{1 + 0.712}{1 - 0.712}\right) = 0.891.$$

This coefficient is approximately normally distributed, with a standard error of

$$s = \frac{1}{\sqrt{N - 3}} = \frac{1}{\sqrt{443 - 3}} = 0.0477.$$

Using this knowledge we can calculate a 95% confidence interval, with the lower bound computed as

$$\hat{z}_L = \hat{z} - 1.96 \times s = 0.891 - 1.96 \times 0.0477 = 0.798,$$

and the upper value computed as

$$\hat{z}_U = \hat{z} + 1.96 \times s = 0.891 + 1.96 \times 0.0477 = 0.984.$$

The conversion of these bounds to a confidence interval around $\hat{\rho}$ can be achieved using Table 7.2. The table contains only positive value of $\hat{\rho}$ and $\hat{z}$. Negative values of $\hat{z}$ can be obtained by reversing the sign. The $\hat{\rho}$ associated with $\hat{z}_L = 0.798$ is 0.66 and the $\hat{\rho}$ associated with $\hat{z}_U = 0.984$ is 0.75. Therefore, the 95% confidence interval is

$$0.66 \leq \hat{\rho} \leq 0.75.[2]$$

Figure 7.7 shows the relationship between the width of the 95% confidence interval and the number of observations used to calculate the estimate of correlation (which we hold constant at 0.712). The picture that emerges is quite clear, the larger the number of observations used to calculate the estimate of correlation, the smaller the confidence interval. Tighter confidence intervals can be obtained by using a larger number of observations.

7.8   HYPOTHESIS TESTS OF THE CORRELATION COEFFICIENT

Correlation coefficients are useful descriptive measures of the strength of linear association between two variables, they can also be used as the basis

Table 7.2 Lookup table for Fisher *z* scores and correlation estimate

Fisher *z* score	Correlation estimate
0	0
0.01	0.01
0.02	0.02
0.03	0.03
0.04	0.04
0.05	0.05
0.0601	0.06
0.0701	0.07
0.0802	0.08
0.0902	0.09
0.1003	0.1
0.1104	0.11
0.1206	0.12
0.1307	0.13
0.1409	0.14
0.1511	0.15
0.1614	0.16
0.1717	0.17
0.182	0.18
0.1923	0.19
0.2027	0.2
0.2132	0.21
0.2237	0.22
0.2342	0.23
0.2448	0.24
0.2554	0.25
0.2661	0.26
0.2769	0.27
0.2877	0.28
0.2986	0.29
0.3095	0.3
0.3205	0.31
0.3316	0.32

Continued

Table 7.2 Continued

Fisher *z* score	Correlation estimate
0.3428	0.33
0.3541	0.34
0.3654	0.35
0.3769	0.36
0.3884	0.37
0.4001	0.38
0.4118	0.39
0.4236	0.4
0.4356	0.41
0.4477	0.42
0.4599	0.43
0.4722	0.44
0.4847	0.45
0.4973	0.46
0.5101	0.47
0.523	0.48
0.5361	0.49
0.5493	0.5
0.5627	0.51
0.5763	0.52
0.5901	0.53
0.6042	0.54
0.6184	0.55
0.6328	0.56
0.6475	0.57
0.6625	0.58
0.6777	0.59
0.6931	0.6
0.7089	0.61
0.725	0.62
0.7414	0.63
0.7582	0.64
0.7753	0.65

Table 7.2 Continued

Fisher *z* score	Correlation estimate
0.7928	0.66
0.8107	0.67
0.8291	0.68
0.848	0.69
0.8673	0.7
0.8872	0.71
0.9076	0.72
0.9287	0.73
0.9505	0.74
0.973	0.75
0.9962	0.76
1.0203	0.77
1.0454	0.78
1.0714	0.79
1.0986	0.8
1.127	0.81
1.1568	0.82
1.1881	0.83
1.2212	0.84
1.2562	0.85
1.2933	0.86
1.3331	0.87
1.3758	0.88
1.4219	0.89
1.4722	0.9
1.5275	0.91
1.589	0.92
1.6584	0.93
1.738	0.94
1.8318	0.95
1.9459	0.96
2.0923	0.97
2.2976	0.98
2.6467	0.99
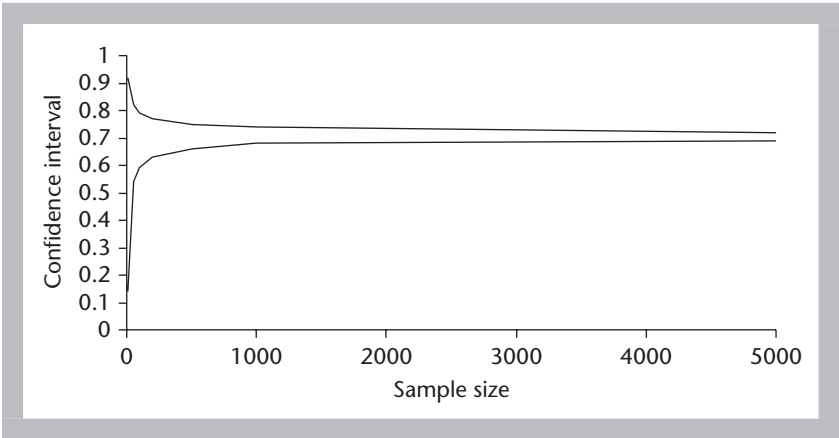
Figure 7.7 Confidence intervals by sample size for a correlation
estimate of 0.712

of a test of the hypothesis of no linear association. To investigate this we
use the function cor.test(x, y, alternative = c("two.sided",
"less", "greater"), method = "Pearson" or "spearman"),
conf.level = confidence level).

**Example 7.5   Hypothesis tests of the correlation between
unemployment and real GNP**

We return to Example 7.2, and investigate whether the previously discussed
measures of correlation are significantly different from zero. We begin with
the product moment coefficient

```
>cor.test(logChange.GNP,logChange.Unemp,method="pearson", conf.le
vel =0.95,alternative="less")
         Pearson's product-moment correlation
data:   logChange.GNP and logChange.Unemp
t = -1.8342, df = 13, p-value = 0.0448
95 percent confidence interval:
 -1.00000000 -0.01416498
sample estimates:
       cor
-0.4534178
```

The function gives the test statistics, an approximate confidence interval,
and the estimate of the product moment coefficient. Since the p value is mar-
ginally less than the level of significance (0.05) we reject the null hypothesis

of zero correlation. The Kendal Tau coefficient is more conclusive

```
>cor.test(logChange.GNP,logChange.Unemp,method="kendal",
conf.level =0.95,alternative="less")
         Kendall's rank correlation tau
data:  logChange.GNP and logChange.Unemp
T = 45, p-value = 0.2475
tau
-0.1428571
```

The Spearman rank provides similar evidence

```
>cor.test(logChange.GNP,logChange.Unemp,method=
"spearman",conf.level =0.95,alternative="less")
         Spearman's rank correlation rho
data:  logChange.GNP and logChange.Unemp
S = 678, p-value = 0.2249
sample estimates:
        rho
-0.2107143
```

The weight of the graphical and correlation evidence indicates no (or a very weak) association between the rate of change in GNP and the rate of change in unemployment.[3]

7.9   COEFFICIENT OF DETERMINATION

The coefficient of determination is the square of the correlation coefficient. Its value, given various values of the correlation coefficient, is calculated below using *R*

```
> correlation =seq(from =-1, to =+1, by= 0.1)
> coef.det=correlation^2
> table=data.frame(correlation,coef.det)
> table
   correlation coef.det
1         -1.0     1.00
2         -0.9     0.81
3         -0.8     0.64
4         -0.7     0.49
5         -0.6     0.36
6         -0.5     0.25
7         -0.4     0.16
8         -0.3     0.09
```

9	-0.2	0.04
10	-0.1	0.01
11	0.0	0.00
12	0.1	0.01
13	0.2	0.04
14	0.3	0.09
15	0.4	0.16
16	0.5	0.25
17	0.6	0.36
18	0.7	0.49
19	0.8	0.64
20	0.9	0.81
21	1.0	1.00

From this output we see the coefficient of determination takes values between 0 and 1. The interpretation of the coefficient of determination is the proportion of variance in one variable, explained by the second variable; For example a correlation between X and Y of ± 0.3 implies that around 9 percent of the variance of X can be explained by Y (or vice versa). A correlation of 0.70 implies a coefficient of determination of 0.49, so that approximately half of the variation in one variable is explained by the other.

7.10 TIME EVOLUTION OF CORRELATION COEFFICIENTS

Estimates of correlations between financial instruments tend to vary over time, as Eydeland and Wolyniec[4] state:

> It is well know [n] that correlations in financial markets tend to be quite unstable. However, the situation in energy markets is much more extreme. Correlations between spot power prices and the corresponding spot fuel prices exhibit instabilities not seen in any other markets.

We see this clearly in Figure 7.8 which shows the rolling 30- and 90-day correlation between Jet fuel spot and 3-month Brent Crude future prices. Correlations can spike at high levels and fall off dramatacally. This volatility is particularily marked in the 30-day rolling estimate. This raises the issue of what is the appropriate window to use? The answer depends on what is causing the variation in correlation. If correlation is deemed to be constant, the variation in Figure 7.8 will be the result of estimation noise. This implies the shorter the window the more volitile will be the estimate of the correlation coefficient, driven by a larger standard error inherent in smaller samples. The solution is to increase the window. However, if the variation is due to changes in the conditional correlation between the insturments, too
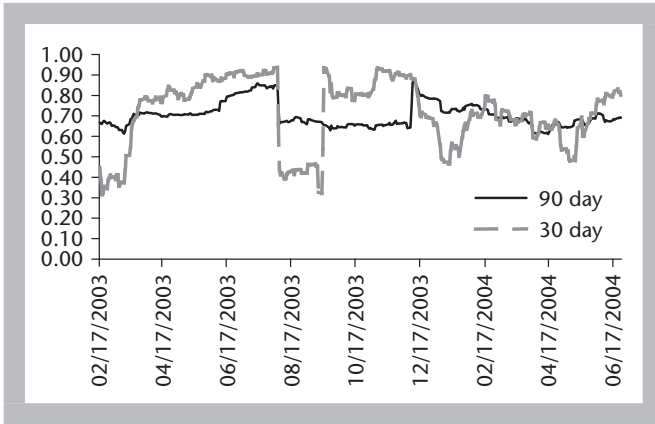
Figure 7.8 Rolling 30- and 90-day correlation between Jet fuel spot and 3-month Brent Crude future prices

large a window may hide important changes in the correlation sturcture. One possible solution is to exponential weight the observations so that the more recent observations have more weight that distant observations.[5] We discuss this idea further in Chapter 12.

Another feature of the time evolution of energy prices is autocorrelation. Autocorrelation (often called serial correlation) is the correlation between a random variable and past values of itself. High levels of autocorrelation indicate that values in the past are associated with present values. Figure 7.9 shows the autocorrelation function (ACF) for the end of month price of Unleaded premium New York gasoline (Non-Oxy). Notice how the correlation between today's price and previous prices declines steadily, eventually becoming negative beyond ten months.

7.11   OTHER MEASURES OF CORRELATION

Whilst Spearman's rank and Pearson coefficients are the most commonly used, there are other measures which can be used when the random variables are not continuous. In this Section we describe three other less frequently seen, but useful measures. They are the point biserial correlation coefficient, Tetrachoric correlation coefficient and Phi coefficient.

7.11.1   Point biserial correlation coefficient

So far we have considered the situation where both variables are continuous. When one of the variables is binary and the other continuous the point biserial correlation coefficient can be used to assess the degree of association. Suppose X is a continuous variable and Y a binary variable taking the
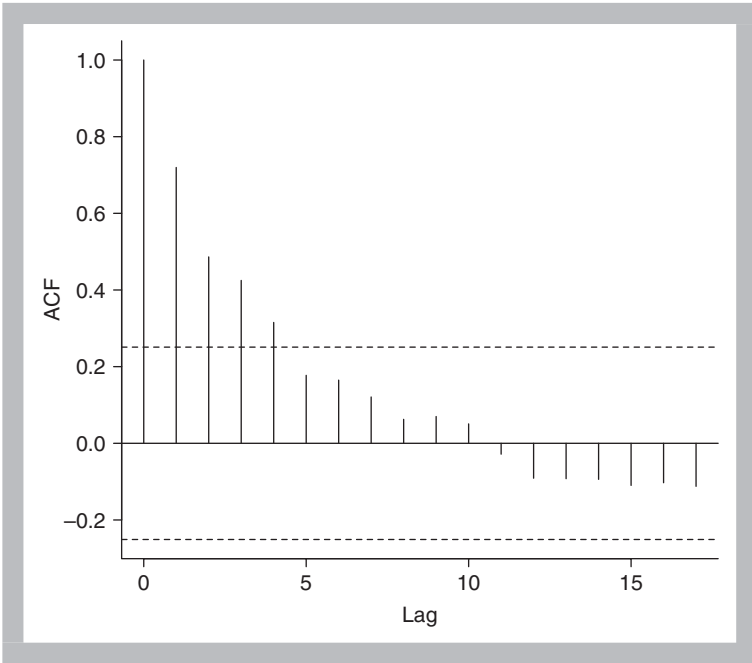
Figure 7.9 Autocorrelation function of the price in Unleaded premium New York gasoline (Non-Oxy)

values 0 and 1, the point biserial correlation is calculated as

$$\rho = \frac{(\bar{X}_1 - \bar{X}_0)\sqrt{p(1-p)}}{S_X},$$

where

$\bar{X}_1$ is the mean of X when $Y = 1$
$\bar{X}_0$ is the mean of X when $Y = 0$
S_X is the sample standard deviation of X
p is the proportion of values where $X = 1$.

A R-function to estimate the biserial correlation coefficient is given below. For the function to work correctly the first column passed to the function should be the continuous variable and the second column the binary variable. Both variables should have the same number of observations

```
bis.corr<-function(contin.var,binary.var)
{
p=sum(binary.var)/length(contin.var)
sdx=sd(contin.var)
1:length(contin.var)
x1.number= sum(binary.var)
```

```
x0.number= length(contin.var)-x1.number
x1=0
x0=0
for(k in 1:length(contin.var))
{
{
    if(binary.var[k]== 1)
           {
           x1= x1+contin.var[k]

           }
else
    {
            x0= x0+contin.var[k]
}
}
}# next for
x1=x1/x1.number
x0=x0/x0.number
  ((x1-x0)*(p*(1-p))^0.5)/sdx
}
```

Example 7.6   Natural gas storage swaps based on Energy
Information Administration data

Consider an over-the-counter swap based on the reported value of US natural gas inventory data supplied by the US Department of Energy's statistical arm (Energy Information Administration). If the swap is a day of release, no revisions, based on the first figure released by the Energy Information Administration, then the risk manager and traders will be interested in the correlation between the gas price returns and the levels of gas storage. Assume

(1) the "seller" of the swap will pay \$$x$ to the energy firm if the figure released by the Energy information Administration is greater than some value θ; and

(2) the energy firm will pay \$$x$ to the "seller" of the swap if the figure released by the Energy Information Administration is less than θ.

Suppose further that we have observed the following returns to the gas price in recent weeks

{11.778, 16.184, 11.711, 16.191, 16.211, 8.045, 5.057, 17.288}

Furthermore, suppose we have also observed over the same period the following sequence when the figure released is greater than θ

$\{0, 1, 0, 1, 1, 1, 0, 1\}$

The point biserial correlation coefficient can be easily calculated in *R*. First we enter the data

```
>PriceReturns=c(11.778,16.184,11.711,16.191,16.211,
8.045,5.057,17.288)
> greaterthan=c(0,1,0,1,1,1,0,1)
```

The variable `greaterthan` is binary taking the value 1 when the figure released by the Energy Information Administration was greater than θ. Since `greaterthan` is dichotomous we calculate the correlation between the two variables using the function `bis.corr()`:

```
> bis.corr(PriceReturns, greaterthan)
[1] 0.5713
```

Thus we have an arrived at a quantitative estimate of the correlation between the two variables.

7.11.2   Tetrachoric correlation coefficient

Tetrachoric correlation measures the degree of association between two binary variables. We shall discuss its calculation with the following illustration. A risk manager of a large financial firm is writing a report about minor breaches in risk limits in all areas of trading across his company. At present some (but not all) of the trading desks have been subjected to an audit, and others have not. The risk manager is interested in measuring the correlation between breaches in trading limits and whether or not a particular trading desk had been audited. The data is presented below in *R* format

```
>breach=c("yes","no","yes","no","no","yes","no","no",
"yes","yes","no","no")
>audit=c("no","no","yes","no","no","yes","no","no",
"yes","yes","yes","yes")
> table(breach,audit)
              Audit
 breach      no  yes
        no   5    2
        yes  1    4
```

Since both `breach` and `audit` are binary the risk manger could calculate the Tetrachoric correlation between them. Unfortunately, the actual formula for the Tetrachoric correlation is complex containing an infinite series of terms. However, the following approximation often works well

$$\hat{\rho}_T = \cos\left[\frac{180°}{\left(1 + \sqrt{bc/ad}\right)}\right],$$

where a, b, c, d refer to the frequencies in a fourfold table in cell 11, 12, 21, and 22 respectively. In terms of the above data $a = 5, b = 2, c = 1$, and $d = 4$, so that we have

$$\hat{\rho}_T = \cos\left[\frac{180°}{\left(1 + \sqrt{2 \times 1/5 \times 4}\right)}\right] = -0.7284.$$

So it appears the correlation is fairly strongly negative. However, the estimate should be viewed with a touch of skepticism because the Tetrachoric correlation coefficient is a only a good estimator of the population correlation when the sample size is reasonably large. In this case, we only have 12 observations.

7.11.3   Phi correlation coefficient

The Phi coefficient is an alternative to the Tetrachoric correlation coefficient for dichotomous data. To estimate Phi we use the formula

$$\hat{\rho}_{Phi} = \frac{(a \times b) - (b \times c)}{\left[(a + b) \times (c + d) \times (b + d)\right]^{0.5}}$$

where a, b, c, d refer to the frequencies in a fourfold table in cell 11, 12, 21, and 22 respectively. It turns out that this is the value calculated by the product moment correlation coefficient when both variables are binary variables. Returning to the data in the illustration, suppose we code yes as a 1 and no as a zero then we would have the two variables

```
breach.binary=c(1,0,1,0,0,1,0,0,1,1,0,0)
```

and

```
audit.binary=c(0,0,1,0,0,1,0,0,1,1,1,1).
```

The correlation between these two variables is therefore

```
> cor(breach.binary,audit.binary)
[1]  0.5070926.
```

Compare this to the value of the Tetrachoric correlation coefficient. Why the big difference between the Tetrachoric estimate and Phi estimate? In estimating Phi the values for the dichotomous observations are arbitrarily assigned to the categories so the sign of Phi is of very little use. It only indicates which diagonal had the greater concentration of scores. However we could, if we wished, contrast the implications of the Tetrachoric estimate and Phi estimate using the coefficient of determination.

7.12   CAUSATION, DEPENDENCE, AND CORRELATION

Correlation may exit between two variables when one of them is related to the other. However, we must always remember that correlation measures the degree of linear strength between two variables. Suppose that a variable $y = \cos(x)$. Let us calculate the correlation between these two variables in R

```
> x=seq(1,10,.001)
> y=cos(x)
> cor(x,y)
[1]  -0.002722781.
```

So although we know the exact functional form of the two variables the estimate of correlation is close to zero. If we were only to look at the correlation coefficient, we would draw the incorrect conclusion of little association between the two variables. However a scatter plot, shown in Figure 7.10, reveals the extent of the non-linearity between the two variables. The important point to note is that any conclusion of no or little linear correlation maybe a consequence of the two variables being related in some non-linear fashion. Such non-linearity can often be revealed by inspecting a scatter plot.

If X causes Y then we know that X is a causal factor of Y. In many cases we are interested in identifying causal factors. For example a natural gas trader would like to identify the casual factors of gas price changes. The trader may observe a high correlation between the change in price of a stock market index or government bonds and the change in price of natural gas. However, a high correlation does not necessarily imply causation. It only suggests that the two variables move together. Just because two variables are highly correlated does not mean that one causes the other. In statistical terms, we say that correlation does not imply causation. Why? First the high correlation may be through chance alone, as appears to be the case between
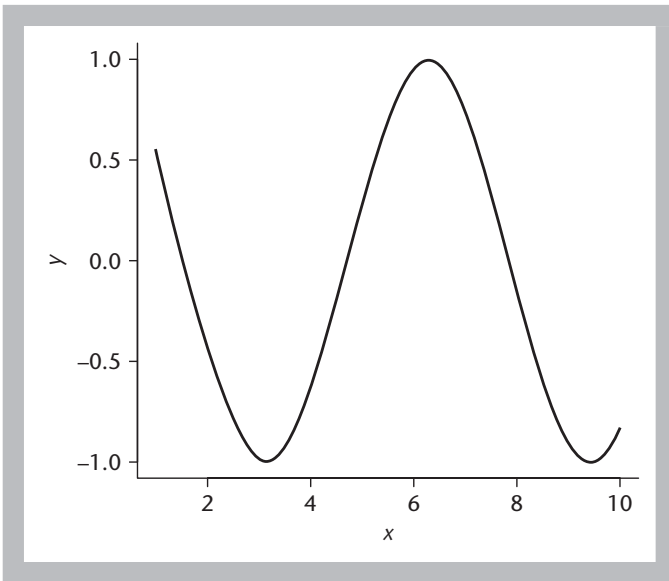
Figure 7.10 Scatter plot of two non-linearity related random variables

the level of GNP and Unemployment. Second, it may be the result of a third variable. As an illustration of this, consider the correlation between height and weight in a random sample of children in the London borough of Ealing. What is the correlation likely to be? It will be high and positive. This does not imply height causes weight or vice versa. In fact the positive correlation between height and weight is a consequence of a third variable, age; older children tend to be taller and heavier than younger children. It is important to remember not to deduce causation from correlation alone. Correlation does not imply causation.

7.13   SUMMARY

A primary source of risk is price fluctuation in energy contracts. Correlation coefficients provide a scale against which the closeness of the relationship between price returns can be measured. For obvious reasons, energy commodity prices are highly correlated. Gasoline prices in various markets will be highly correlated with each other, as will prices of very similar commodities such as heating oil and diesel. There are many estimators of correlation, they usually provide information about two aspects of the relationship between variables – its strength, and its direction. In using a measure of correlation it is always advisable to plot the data to ensure that the inherent assumption of linearity is valid.

7.14 FURTHER READING

Details of a number of correlation measures alongside easy-to-use implementations in the Excel spreadsheet package are given Lewis (2004). Further analysis of the role of correlation and risk management can be seen in the web paper by Embrechts *et al.* entitled "Correlation and dependence in risk management :properties and pitfalls." It is freely available for download from at: www.risklab.ch/ftp/papers/CorrelationPitfalls.pdf

Further examples and the debate surrounding the role of correlation in hedging can be found in Johnson (1960), Ederington (1979), Anderson and Danthine (1981), Dale (1981), Horsewood (1997), Bullimore (2000), Co (2000) and Considine and Heo (2000). Other relevant articles with a slightly more theoretical/statistical focus include Galambos (1987), Harlow (1991), Joe (1997), and Albers (1999).

Albers, W. (1999) "Stop-loss premiums under dependence," *Insurance: Mathematics and Economics*, 24, 173–85.

Anderson, R. W. and Danthine, J. P. (1981) "Cross Hedging," *Journal of Political Economy*, 89 (6) 1182–96.

Bullimore J. (2000) "Jet Fuel Price Risk Management," *Swiss Derivatives Review*, Autumn, 16, 20.

Co, T. (2000) "High Prices Fuel Airlines' Anxiety," *Asia Risk*, November, Special 5th Anniversary Issue, 6.

Considine, T. J. and Heo, E. (2000) "Price and Inventory Dynamics in Petroleum Product Markets," *Energy Economics*, 22, 527–47.

Dale, C. (1981) "The Hedging Effectiveness of Currency Futures Markets," *The Journal of Futures Markets*, 1(1) 77–88.

Ederington, L. H. (1979) "The Hedging Performance of the New Futures Market," *The Journal of Finance*, 34(1) 157–70.

Galambos, J. (1987) *The Asymptotic Theory of Extreme Order Statistics*, Kreiger Publishing Co., Melbourne, FL.

Harlow, W. (1991) "Asset Allocation in a Downside-risk Framework," *Financial Analysts Journal*, 47(5) 28–40.

Horsewood, R. (1997) "Options and Oils," *Asia Risk*, August 1997.

Johnson, L. L. (1960) "The Theory of Hedging and Speculation in Commodity Futures," *Review of Economic Studies*, 27(3) 139–51.

Joe, H. (1997) "*Multivariate Models and Dependence Concepts,* Chapman & Hall, London.

Lewis, Nigel Da Costa (2004) *Operational Risk with Excel and VBA: Applied Statistical Methods for Risk Management*, John Wiley & Sons, Inc., New York.

7.15 REVIEW QUESTIONS

1 What are the restrictions on the use of the product moment correlation coefficient?

2 Consider the correlation between the price of petrol/gasoline and the number of divorces in the USA.
   (a) What would you expect the correlation to be (high/low/negative/positive) and why?
   (b) What transformation/changes to the data, if any, would you make prior to estimating the correlation coefficient between the price of petrol and the number of divorces.

A Primer in Applied Regression Analysis

Regression modeling lies at the heart of modern statistical analysis. It also occupies a key role in much of the analysis carried out in quantitative finance. Given its importance and frequency of use, this chapter provides a hands-on introduction to applied regression modeling. The emphasis is on using R to analyze some simple data sets contained in the R package. The objective is to give you a feel for regression techniques using simple (nonenergy) examples before we move onto discuss more energy-specific applications of regression in the remaining chapters of this book. The emphasis of this chapter is therefore on you the reader becoming comfortable with the ideas surrounding regression and replicating for yourself the examples given in *R*.

8.1   THE SIMPLE LINEAR REGRESSION MODEL

It is often the case that we wish to describe the relation between a set of variables and use this relationship to predict the value of one variable, known as the dependent variable, when we only know the values of the other variables, known as the independent variables. For example , we might be interested in the relationship between the natural gas city-gate price in New York City and the Henry Hub price. Correlation is of little use in this situation as it only describes the degree of association between the variables. How should we characterize the relationship? Regression modeling is one way we can specify and test a particular postulated relationship. We begin by postulating a regression function and then make estimates of model parameters given values of the dependent and independent variables.

The simplest functional form for relating a dependent variable to one or more independent variable(s) is a straight line. We might therefore speculate that the functional form of the relationship between the variable we wish to predict and the other variable(s) is linear. This is the basis of linear regression.[1] Although we cannot expect the postulated linear relationship to be perfect, we might seek to establish its approximate validity. Once a linear relationship is established, knowledge of the independent variable(s) can be used to inform us about plausible values of the dependent variable. Suppose we observe a sample of N pairs of observations $\{(y_1, x_1), (y_2, x_2), ..., (y_N, x_N)\}$ on two continuous variables X and Y. We wish to describe the relationship between them, and thus be able to predict the value of Y given knowledge of X. Clearly the correlation between the two variables is of little use as it only indicates the strength of the linear association. Our interest lies in using X to help explain Y. Since X is being used to explain Y it is known as the *independent* or *explanatory* variable, and Y is known as the *dependent* variable. One way of describing the relationship between the realized observations on $X\{x_1, ..., x_N\}$ and $Y\{y_1, ..., y_N\}$ is to assume that it takes a linear form $y_i = \alpha + \beta x_i$.

This is a straight line with intercept α and slope equal to β. When $\beta > 0$ the line has positive slope and a negative slope when for $\beta < 0$. Since we cannot expect this relationship to hold exactly for all N of our paired observations, we include the error term ε_i and write the simple regression equation as

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

The character ε_i is known as the residual and measures the error between the observed y_i and the value that the linear function implies y_i should be. For example, if the linear approximation is

$$y_i = 1 + 2x_i,$$

and we observe $y_1 = 3$ and $x_i = 1$ then $\varepsilon_1 = y_1 - (\alpha + \beta x_1) = 0$. However, if $y_2 = 2$ and $x_2 = 3$ then $\varepsilon_2 = 2 - (1 + (2 \times 3)) = -5$. We will therefore obtain N measurements on ε_i. Since ε is a random variable it will have an underlying probability distribution. We shall assume ε_i to be an independent identically normally distributed random variable with mean zero and variance equal to σ^2.

8.2   EXPECTATION AND REGRESSION

The expected value of a random variable is a measure of its average or typical value. The expected value, denoted by $E[X]$, of a discrete random variable

X is calculated as

$$E[X] = \sum_x x \times \text{probability}(x)$$

Suppose after a few drinks, despite being a risk averse risk manager, you feel lady luck is on your side. As a one-off event you decide to buy a lottery ticket. The two possible outcomes of the random variable $X = lottery\ outcome$ are win and lose. If you win the random variable $X = 1$, and for failure $X = 0$. This random variable is known as a Bernoulli random variable. Let the probability of winning the lottery be $\text{Prob}(X = 1) = p$. What is the expectation of this random variable?

$$E[X] = \sum_x xp(x) = 0(1 \times -p) + (1 \times p) = p.$$

Of course p tends to be rather small, for the United Kingdom national lottery p is approximately equal to $1/14,000,000$.

If the random variable X is continuous $E[X]$ is calculated by

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx.$$

Let us consider the calculation of the expected value for a continuous random variable X, which takes values between 0 and 1 with equal probability. Such a random variable is said to be uniformly distributed. The probability density function of this random variable is

$f(x) = 1$ if X lies between 0 and 1 otherwise $f(x) = 0$.

What is the expected value of X?

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx = \int_0^1 x\, dx = 1/2.$$

Finally we note, if Y is some function of X, denoted by $Y = G(X)$, then

1. if X is discrete: $E[Y] = \sum_x G(x)p(x)$
2. if X is continuous: $E[Y] = \int_{-\infty}^{\infty} G(x)f(x)dx$

8.2.1   Calculating expected values in *R*

To illustrate the calculation of an expected value,[2] consider the number of risk analysts available to monitor trading activity on a particular day.

The head of risk management has been asked to report on the expected level of staffing. He knows that on any day there is always at least one risk analyst available for duty and at most nine. Therefore the discrete random variable X can take on values between 1 to 9. Suppose he calculates the probability of each specific value as

```
>Probability=c("1"=0.301,"2"=0.176,"3"=0.125,"4"=0.097,"5"=0.079
,"6"=0.067,"7"=0.058,"8"=0.051,"9"=0.046)

> Probability

1       2     3     4     5     6     7     8     9
0.301 0.176 0.125 0.097 0.079 0.067 0.058 0.051 0.046
```

We see that the probability that $X = 1$ is 0.301 and the probability $X = 9$ is 0.046. In other words there is only a 4.6% likelihood of all nine staff being available on any one day. To calculate the expected value we create a vector `Staff` and multiply it by `Probability` storing the result in a vector `Result`. The sum of `Result` gives the expected value

```
> Staff =1:9
> Result =Staff*Probability
> Result
  1       2     3     4     5     6     7     8     9
  0.301 0.352 0.375 0.388 0.395 0.402 0.406 0.408 0.414
> ExpectedStaff=sum(Result)
> ExpectedStaff
[1] 3.441
```

Thus $E[X] = 3.44$. This informs the head of risk that the average or center of mass of the distribution lies between 3 and 4. However, as the actual number of staff on duty can only take integer values the expected value will never be observed. In reporting his findings the head of risk is likely to say that he expects the number of personnel available on a particular day to be around 3 or 4.

8.2.2   Law of large numbers and expected values

The law of large numbers states if we perform N independent and identical random experiments on a random variable X, as N goes to infinity the average of the N outcomes approaches the expected value $E(X)$. Let us use R to

investigate the law of large numbers. The function `runif(N)` generates *N* random variables from a uniform distribution. We already know the expected value of this distribution is equal to 0.5. We use the law of large numbers to estimate it by generating random samples of various sizes

```
> x1=runif(1)
> x2=mean(runif(10))
> x3=mean(runif(100))
> x4=mean(runif(1000))
> x5=mean(runif(10000))

> X=cbind(x1,x2,x3,x4,x5)
    round(X,5)
         x1      x2       x3       x4       x5
[1,] 0.41596   0.53344 0.50503 0.49965 0.50045
```

As the sample gets larger the average calculated using `mean()`, approaches 0.5. Now we can see the expected value is essentially the long run probability weighted average over a large number of trials. That is to say if we perform a random experiment on many occasions and take the probability weighted average, this will be the expected value. Since it is an average, its value may never actually be observed as we saw in Section 8.2.1.

8.2.3   Rules of expected values

There are a number of rules of expectation which are useful in practical applications:

Rule 1: The expectation of a constant, C, is the constant,

$$E[C] = C.$$

Rule 2: If X is a random variable and C a constant then,

$$E[XC] = CE[X]$$

Rule 3: The expected value of the sum of two random variables X and Y is the sum of their expected values

$$E[X + Y] = E[X] + E[Y].$$

Rule 4: Given two random variables X and Y, the conditional expectation of X given Y is denoted by $E[X|Y]$. If X is continuous it is defined by

$$E[X|Y] = \int_{-\infty}^{\infty} xf(x|y)dx.$$

If X is discrete it is defined by

$$E[X|Y] = \sum_x xp(x|y).$$

We can interpret linear regression as a conditional expectation where $E(Y|X) = \alpha + \beta X$ and α the intercept tells us the value of Y that is expected when $X = 0$. The slope parameter β measures the relationship between X and Y. It is interpreted as the expected or average change in Y for a 1-unit change in X. For example, if we estimate a regression and find $E(Y|X) = 1 + 2X$ a 1-unit change in X is expected to lead to a 2-unit change in Y.

8.3   PARAMETER ESTIMATION

What value should we choose for the regression model parameters α and β? We have already discussed methods of estimation such as ordinary least squares (OLS) and maximum likelihood, can we use these techniques here to obtain an estimator? The answer is yes. We can use the method of maximum likelihood or the method of OLS to obtain estimators of the parameters. Given the sample mean of Y and X, denoted by $\bar{X}$ and $\bar{Y}$ respectively, OLS estimators of the parameters are given by $\hat{\alpha} = \bar{Y} - \beta\bar{X}$ for the intercept parameter, and for the slope parameter

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{N}(x_i - \bar{X})^2}$$

We can also estimate the intercept and slope parameters using maximum likelihood. The individual y_is are assumed to be independent, identically normally distributed with mean population mean μ and population variance σ^2. The log likelihood equation is given by

$$\log L\left(\alpha, \beta, \sigma, \mu | \left\{(y_1, x_i), ..., (y_n, x_n)\right\}\right)$$
$$= -\frac{1}{2}N\log 2\pi - \frac{1}{2}N\log \sigma^2 - \frac{1}{2}\sum_{i=1}^{N}\frac{(y_i - \alpha - \beta x_i)^2}{\sigma^2}$$

which on solving yields

$$\hat{\alpha} = \overline{Y} - \beta\overline{X}, \quad \hat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{N}(x_i - \overline{X})^2}.$$

8.3.1   Parameter estimation using the *R* statistical package

Estimation of linear regression parameters is carried out using the function `lm(regression model, data)`. The model is specified symbolically using the notational form `dependent variable ~ independent variable`. In general we will estimate the parameters of our postulated model using this function, assigning the values to an object of our choice.

The prepackaged *R* data frame[2] `cars` gives the speed of cars (`speed`) recorded in miles per hour and the distances (`dist`) taken to stop, in feet, of motorized vehicles recorded in the 1920s. The product moment correlation between the two variables is 0.81 and the rank correlation coefficient is 0.83. The data are shown in the scatter plot of Figure 8.1. As the data appear reasonably linear we postulate the following linear relationship between the distance it takes to stop (dependent variable) and the speed of the car (independent variable): Distance= $\beta \times$ Speed + Error. Notice in this model
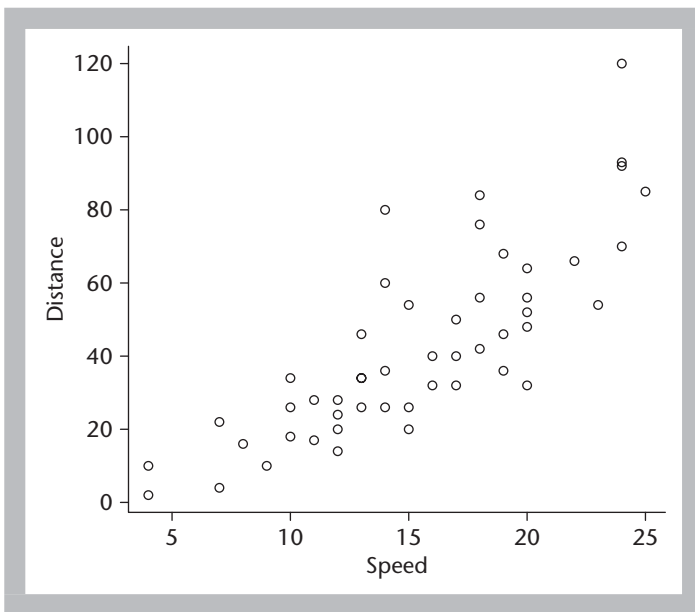

Figure 8.1 Scatter plot of the speed of cars (miles per hour) and the distances taken to stop (feet)

we have set the intercept term α to equal zero, but given the upward slope shown by the scatterplot we would expect $\beta > 0$.

We can estimate the model in *R* as follows

```
> data(cars)
> names(cars)
[1] "speed" "dist"
> car.reg=lm(dist~speed-1,data=cars)
> car.reg
Call:
lm(formula = dist ~ speed - 1, data = cars)
Coefficients:
speed
 2.91
```

Observe the output gives both the model `lm(dist~speed-1,data=cars)` and the coefficient estimate (β) on the independent variable `speed` which is equal to 2.91.[3] We can see the elements of `car.reg` by using the function `names()`

```
names(car.reg)
  [1] "coefficients"  "residuals" "effects" "rank"
  [5] "fitted.values" "assign"    "qr"
"df.residual"
 [9]  "xlevels"       "call"       "terms"    "model"
```

This informs us that we could also use the `coefficients()` (which we shorten to `coef()`) to give us the estimated parameter values

```
> options(digits=3)
> coef(car.reg)
speed
 2.91
```

However, notice that `coef()` does not return details of the model. In the above `lm()` function we used "−1" to indicate that the model should be estimated without an intercept term. If we wished to include an intercept term we could use

```
> car.reg=lm(dist~speed+1,data=cars)
> car.reg
Call:
lm(formula = dist ~ speed + 1, data=cars)
```

```
Coefficients:
(Intercept)        speed
     -17.58        3.93
```

In this case the coefficient on speed increases to 3.93 and we see the intercept takes a value of –17.58. The default of `lm()` includes the intercept, so we do not really need to specify "+1"

```
> car.reg=lm(dist~speed,data=cars)
> car.reg
Call:
lm(formula = dist ~ speed, data = cars)
Coefficients:
(Intercept)          speed
     -17.58          3.93
```

8.3.2   Hedging energy price risk using simple linear regression

In Chapter 7 we saw, given the futures F and spot S, that the minimum variance hedge ratio was given by

$$h = \frac{\text{Covariance}(F, S)}{\sigma_F^2}$$

Since we can interpret linear regression as a conditional expectation where the slope parameter β measures the relationship between the two variables we can use the simple linear regression model to obtain an estimate of h. To do this we would specify the regression model

$$\Delta S_t = \alpha + \beta \Delta F_t + \varepsilon_t$$

where ΔS_t and ΔF_t are the spot and futures returns for period t. The OLS or maximum likelihood estimator for β provides an estimate for the minimum variance hedge ratio h.

Example 8.1   Hedging the Jet kerosene spot using the Brent Crude forwards

The basic principles of hedging can be used for energy commodities for which no futures or forward contracts exists, provided similar commodities are traded. As an example we consider hedging Jet kerosene with Brent Crude oil forward contracts. The correlations between the log return of the spot price of Jet kerosene and forward returns for the various contracts on Brent Crude oil are presented in Table 8.1. The liner regression estimates of the hedge ratio are given in Table 8.2, in all cases the intercept is estimated

Table 8.1 Correlation between the log returns of Jet kerosene and Brent Crude forwards contracts (October 2002 to June 2004)

Brent Crude	Jet kerosene
1-month	0.728
2-month	0.729
3-month	0.712

Table 8.2 Simple linear regression estimates of minimum variance hedge ratio

Regression parameter	Forwards contract		
	1-month	2-month	3-month
β	0.970	1.030	1.044
α	0.000	0.000	0.000

at zero, with the slope parameter close to 1. Since we can interpret linear regression as a conditional expectation where $E(\Delta S_t|\Delta F_t) = \alpha + \beta \Delta F_t$. We see immediately that the slope parameter β provides a natural estimate for h because it measures the relationship between ΔS_t and ΔF_t. For example, for the 1-month forward contract we find

$$E(\Delta S_t|\Delta F_t) = 0.970 \ \Delta F_t.$$

Therefore, a 1-unit change in ΔF_t is expected to lead to a 0.970 unit change in ΔS_t.

8.4 ASSESSING THE SIMPLE LINEAR REGRESSION MODEL

Given the estimates of the model parameters, we will be interested in assessing how well the model describes the sample data. Related questions include:

■ How well does the model explain the data?

■ Are the values of the estimated slope and intercept significantly different from zero?

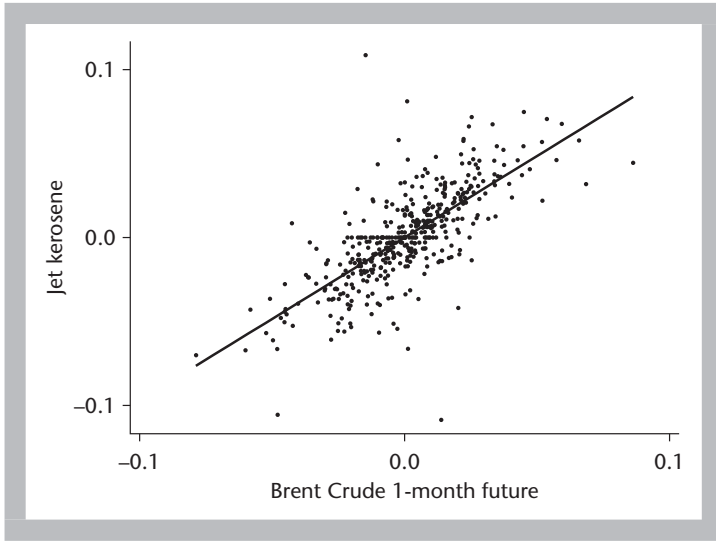■ Are the estimates equal to some postulated values?

Figure 8.2 Scatter plot of the log returns of Jet kerosene on 1-month Brent Crude future with regression line superimposed

We can investigate these issues with a combination of hypothesis tests, regression plots, and the coefficient of determination.

8.4.1   Regression plots

A scatter plot with the estimated regression line superimposed (know as a regression plot) Provides the simplest way to assess the fit of the estimated regression. Figure 8.2 shows a scatter plot of the log returns of Jet kerosene and the 1-month Brent Crude future with the estimated regression line superimposed. Since the plotted points do not line exactly along this line, the estimated regression education does not fully explain the data. Nevertheless, as a first approximation for hedging, the model looks adequate.

8.4.2   *t*-test of a regression coefficient

In Section 8.3.1 we estimated two simple regression models, with and without the intercept term. Which should we choose? One way we can make a decision is to investigate the statistical significance of the intercept term via the null hypothesis test that it is zero. If we reject the null hypothesis we include the intercept term in the model. The test statistic in this case is

$$t_{\alpha/2} = \frac{\hat{\alpha}}{\text{s.e.}(\hat{\alpha})}$$

where s.e.($\hat{\alpha}$) is the standard error of the estimate of the intercept. What precisely is the standard error of the estimate? As the sample from which we derive the parameter estimates is a random sample, the estimates themselves, which are a function of the data, are also random variables. Their value will change from sample to sample. We estimate the variation in the parameter estimates using the standard deviation of the estimate, more frequently called the standard error of the estimate. Furthermore, the sampling distribution of the above test statistic is the Student t distribution with N-2 degrees of freedom. We can also use the above test statistic on the slope parameter, in which case we would specify

$$t_{\alpha/2} = \frac{\hat{\beta}}{\text{s.e.}(\hat{\beta})}.$$

We can even, if we wish, specify a null hypothesis other than zero, say equal to β_0

$$t_{\alpha/2} = \frac{\hat{\beta} - \beta_0}{\text{s.e.}(\hat{\beta})}.$$

Furthermore, given the level of significance equal to α, we can construct a $(1-\alpha)\%$ confidence interval around the estimates via

$$\hat{\beta} \pm \text{s.e.}(\hat{\beta}) \times t_{\alpha/2},$$

and for the intercept

$$\hat{\alpha} \pm \text{s.e.}(\hat{\alpha}) \times t_{\alpha/2}.$$

8.4.3   Regression plots and the *t*-test of a regression coefficient using *R*

Returning to the `cars` data frame[2] we can use R to impose a regression line on a simple scatter plot as follows

```
>plot(cars$speed,cars$dist,xlab="speed",
  ylab="distance")
>abline(car.reg)
```

The resultant plot is shown in Figure 8.3. The function `abline()` adds the estimated regression line to the plot.
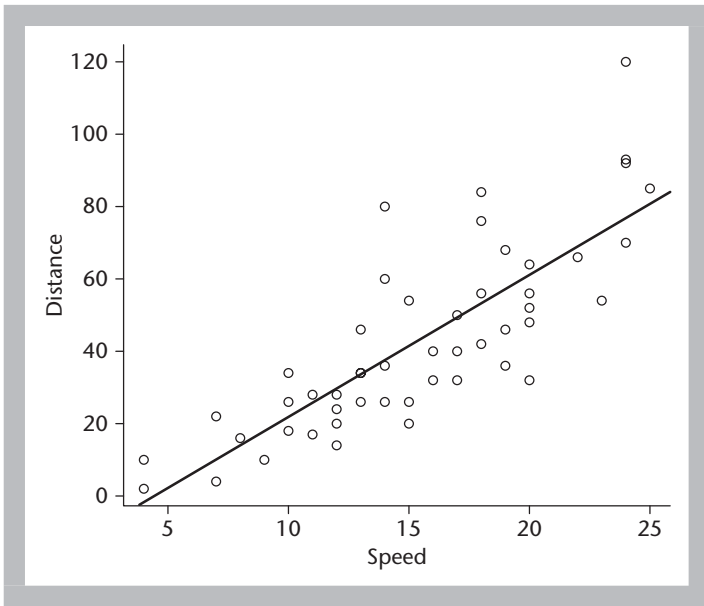
Figure 8.3 Estimated regression line and scatter plot of the speed of cars and distance taken to stop

Continuing with the cars data frame, we can obtain estimates of the standard error of the parameters and *p*-values using the coef() function:

```
> options(digits=4)
> car.reg1 <- summary(car.reg <- lm(dist ~ speed,
  data=cars))
> coef(car.reg1)
            Estimate Std. Error  t value Pr(> |t|)
(Intercept) -17.579  6.7584      -2.601  1.232e-02
speed         3.932  0.4155       9.464  1.490e-12
```

We see the standard error of the estimate of the slope coefficient is 0.4155. The *p*-values is also reported, at the 5% significance level its value is less than 0.05, and we reject the null hypothesis that the coefficient on speed is equal to zero. We draw the same conclusion about the intercept term. A 95% confidence interval around the estimates can be obtained using the function confint(regression object, confidencelevel):

```
> confint(car.reg,level=0.95)
              2.5 %              97.5 %
(Intercept) -31.168            -3.990
speed         3.097             4.768
```

8.4.4   Coefficient of determination

The coefficient of determination, also known as R squared, is frequently used to assess how well the simple linear regression model fits the data. We saw in Chapter 5 that the coefficient of determination is the square of the correlation coefficient between two variables X and Y. If there is no linear relationship between X and Y the correlation coefficient is equal to zero and so therefore is the coefficient of determination. Moreover, since the coefficient of determination is the square of the correlation coefficient it lies between zero (no linear relationship) and one (a perfect linear relationship).

We can interpret the coefficient of determination as the proportion of the variation in the dependent variable Y explained by the linear regression. To see this, recall our parameter estimates were $\hat{\alpha}$ and $\hat{\beta}$, and our estimated linear regression equation was

$$y_i = \hat{\alpha} + \hat{\beta}x_i + \varepsilon_i.$$

The estimated value of y_i can also be written as

$$y_i = \hat{y}_i + \varepsilon_i,$$

where $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ and is the expected value for y_i. Squaring both sides and taking the sum we see that

$$\sum_{i=1}^{N} y_i^2 = \sum_{i=1}^{N} (\hat{y}_i + \varepsilon_i)^2 = \sum_{i=1}^{N} (\hat{y}_i^2 + 2\varepsilon_i\hat{y}_i + \varepsilon_i^2)$$

$$= \sum_{i=1}^{N} \hat{y}_i^2 + \sum_{i=1}^{N} \varepsilon_i^2, \quad \text{as } \sum_{i=1}^{N} 2\varepsilon_i\hat{y}_i = 0$$

since

$$\sum_{i=1}^{N} 2\varepsilon_i\hat{y}_i = 2\sum_{i=1}^{N} \varepsilon_i\hat{y}_i = 0.$$

We have assumed previously that the mean of the residual is zero. Using this assumption and expressing the previous equation in deviation form we obtain

$$\sum_{i=1}^{N} (y_i - \overline{Y})^2 = \sum_{i=1}^{N} (\hat{y}_i - \overline{Y})^2 + \sum_{i=1}^{N} (y_i - \hat{y})^2.$$

This is very similar to the formula for sample variance

$$\text{Var}(Y) = \sum_{i=1}^{N} \frac{(Y_i - \overline{Y})^2}{N - 1}.$$

Now we note the following:

- $\sum_{i=1}^{N}(y_i - \overline{y})^2$ is known as the Total Sum of Squares (TSS)
- $\sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2$ is known as the Explained Sum of Squares (ESS)
- $\sum_{i=1}^{N}(y_i - \hat{y})^2$ is known as the Residual Sum of Squares (RSS).

The difference between TSS and RSS represents the improvement obtained by using the independent variable X to explain Y. This difference is ESS. We see that TSS $=$ ESS $+$ RSS, which is a measure of the total variation in Y explained by both the model and residual.

The coefficient of determination, which we denote by R^2, can be calculated by taking the ratio of the explained variance to the total variance

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

Provided an intercept term is included in the regression, R^2 is simply the proportion of total variation in Y explained by regression model. For an ill-fitting model we see that RSS is large and ESS is small and consequently R^2 will be small. For a well-fitting model ESS is large and RSS small and therefore R^2 will be large. Now we can see:

- If there is no correlation between X and Y then ESS $=0$ and therefore $R^2 = 0$.

- If the correlation between X and Y equals ± 1 then RSS $=0$ and therefore ESS $=$ TSS and $R^2 = 1$.

Example 8.2 R^2 and the minimum variance hedge ratio

In Example 8.1 we estimated the minimum variance hedge ratio using simple linear regression. Now we can see that the higher the coefficient of determination, the greater the effectiveness of the minimum variance hedge. The R^2 statistics for each of the models are given in Table 8.3. We saw in Chapter 5 that the coefficient of determination can also be seen as the square of the correlation coefficient. Table 8.4 presents the values using the Pearson and Spearman rank coefficients. From this table we see first that both coefficients yield similar estimates. Furthermore, the value based

Table 8.3 R-squared statistics for linear regression estimates of minimum variance hedge ratio of the log returns of Jet kerosene on various Brent Crude forwards contracts

	Forwards contract		
	1-month	**2-month**	**3-month**
R^2	52.9%	53.2%	50.7%

Table 8.4 Coefficient of determination between the log returns of Jet kerosene and various Brent Crude forwards contracts calculated using the square of the Pearson and Spearman correlation coefficients

Brent Crude	Pearson (%)	Spearman (%)
1-month	52.9	56.4
2-month	53.2	56.5
3-month	50.7	54.3

on the Pearson coefficient is exactly the same as the value derived from the regression equation. The coefficient of determination in simple linear regression is essentially the square of the Pearson correlation coefficient.

8.4.5 Correlation and coefficient of determination using *R*

Returning to the `cars` data frame[2] we observe

```
> cor(cars["speed"],cars["dist"])
        dist
speed 0.807
> (cor(cars["speed"],cars["dist"]))^2
        dist
speed 0.651
```

The correlation between the variables is high at approximately 81% and the proportion of variation explained by the regression model is around 65%.

We can also use the relationship between TSS, RSS and ESS to directly estimate R^2

```
> y = cars[``dist'']
> y.hat=fitted.values(car.reg)
> y.bar=mean(cars[``dist''])
> ESS=sum((y.hat-y.bar)^2)
> ESS
[1] 21185
> RSS =sum(( y-y.hat)^2)
> RSS
[1] 11354
> TSS=ESS+RSS
> TSS
[1] 32539
```

We saw previously that the coefficient of determination can be calculated as $R^2 = $ ESS/TSS or $R^2 = 1 - $ (RSS/TSS)

```
> R.square = ESS/TSS
> R.square
[1] 0.651
> R.square =1-(RSS/TSS)
> R.square
[1] 0.651
```

Both formulas give identical results and match the value calculated using the square of the sample correlation coefficient. Another direct approach for calculating ESS and RSS in *R* is to use the function anova(regression object)

```
> anova(car.reg)
Analysis of Variance Table
Response: dist
            Df    Sum Sq  Mean Sq    F value Pr(>F)
speed       1 21185      21185    89.6            1.5e-12
***
Residuals 48 11354               237
---
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05`.' 0.1 ` '1
```

Looking at the values under the "Sum Sq" column we see the function ESS = 21185 and RSS = 11354, exactly the same values as calculated previously.

8.5   SUMMARY

Statistical analysis is often focused on questions involving the relationship between one or more variables that are thought as being dependent on some other independent variables. These questions can be addressed using regression modeling. Regression techniques allow us to infer more than possible by measuring the correlation between the dependent variable and independent variables. In this sense it is more flexible and can answer a much wider range of questions. Linear regression can be estimated using `lm()` whilst a broad range of models from the class of generalized linear models (including linear regression) can be estimated via the `glm()` function.

8.6   FURTHER READING

There is a wide array of texts covering regression modeling. For a solid introduction and discussion of the key issues see Doran (1989), Weisberg (1985), or Neter *et al.* (1996). Applications in risk management alongside Excel spreadsheets and Visual Basic source code can be found in Lewis (2004).

Doran, H. E. (1989) *Applied Regression Analysis in Econometrics*, Marcel Dekker, Inc., New York.
Lewis, Nigel Da Costa (2004) *Operational Risk with Excel and VBA: Applied Statistical Methods for Risk Management*, John Wiley & Sons, Inc., New York.
Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996) *Applied Linear Regression Models* (3rd edn), Richard D. Irwin, Inc., Chicago, IL.
Weisberg, S. (1985) *Applied Linear Regression*, John Wiley and Sons, New York.

8.7   REVIEW QUESTIONS

1 Explain the difference between a dependent and an independent variable.

2 What assumptions do we make about the error term in simple linear regression?

3 Re-estimate Example 8.1 using `log(dist)` as the dependent variable and `log(speed)` as the independent variable.
   (a) Report the summary statistics of this regression.
   (b) Comment on the statistical significance of the coefficients (assume the level of significance is 10%)
   (c) How well does the model fit the data relative to Example 8.1?
   (d) Calculate the Pearson correlation coefficient between `log(dist)` and `log(speed)`.
   (e) Now standardize both `log(dist)` and `log(speed)` and re-run the regression – Comment on your findings.

4 What is the relationship between the product moment correlation coefficient and R-squared?

Multiple Regression and Prediction

In this chapter we extend the simple linear regression model into the multiple linear regression model. Multiple regression is useful when we can expect more than one independent variable to influence the dependent variable. It allows us to explore the relationship between several independent and a single dependent variable. We also discuss multivariate regression which arises when we have several dependent variables dependent on the same (or some subset) independent variables.

9.1   THE MULTIPLE REGRESSION MODEL

The multiple linear regression model takes the form

$$y_t = \alpha + \beta_1 x_t^1 + \beta_2 x_t^2 + \cdots + \beta_k x_t^k + \varepsilon_t,$$

where Y is the dependent variable, $\{X^1, X^2, ..., X^k\}$ are k independent variables and ε is the residual. As a simple illustration, consider the situation where we are interested in describing important factors that determine the monthly change in price in the natural gas city-gate price in New York City. Clearly the weather will have an impact, we can measure this using the notion of heating degree days. We might also expect the Henry Hub price, being the largest centralized point for natural gas spot and futures trading in the USA, will also play a role. Our multiple regression would take the form

$$\Delta y_t = \alpha + \beta_1 \Delta x_t^1 + \beta_2 \Delta x_t^2 + \varepsilon_t,$$

where Δy_t, Δx_t^1 and Δx_t^2, denote the monthly log return of the New York city-gate price, Henry Hub price and change in the heating degrees respectively.

9.2　ASSESSING THE MULTIPLE REGRESSION MODEL

As with linear regression, we will need to be able to assess how well the proposed model fits the data. One approach is to carry out a t test on each of the coefficients on the independent variables and the intercept coefficient. A statistically significant result would indicate that the independent variable (or intercept) has an important role to play in explaining the variation in the dependent variable. Two additional approaches we could use are the F-test of joint significance and the adjusted R-squared statistic. We discuss both of these options below.

9.2.1　The *F*-test

The F-test of regression coefficients is a joint test of the null hypothesis that none of the explanatory variables have any effect on the dependent variable. Provided the regression model has an intercept, the test statistic is calculated using

$$F_{k,\,N-k-1} = \frac{[(\text{TSS} - \text{RSS})/k]}{[\text{RSS}/(N - k - 1)]}$$

This test statistic has a F distribution with k and $n-k-1$ degrees of freedom. Rejection of the null hypothesis implies at least one of the coefficients on the explanatory variables is not equal to zero.

9.2.2　Adjusted *R²*

The adequacy of the fit of the multiple regression model cannot necessarily be assessed using R^2. This is because it can be inflated towards its maximum value of 1 simply by adding more independent variables to the regression equation. Instead, the adjusted coefficient of determination is often reported. It takes into account the number of explanatory variables in the model

$$\text{Adjusted } R^2 = 1 - \left[\frac{\text{RSS}/(N - k)}{\text{TSS}/(N - 1)} \right].$$

If we only have one independent variable in our model (so that we have a simple linear regression) then $k = 1$ and we see that

$$\text{Adjusted } R^2 = 1 - \left[\frac{\text{RSS}/(N-k)}{\text{TSS}/(N-1)}\right] = 1 - \frac{\text{RSS}}{\text{TSS}} = R^2.$$

9.3   PREDICTION

Regression is used both for descriptive and predictive purposes. When used for prediction our interest lies in possible future values of the dependent variable given the independent variables. In simple regression to obtain a predicted value $\hat{y}_{i+1}$ we use

$$\hat{y}_{i+1} = \hat{\alpha} + \hat{\beta} x_{i+1}$$

To illustrate this idea, suppose we estimate $\hat{\alpha} = 1$ and $\hat{\beta} = 3$. We then observe $x_{i+1} = 2$, the predicted value of Y is therefore $\hat{y}_{i+1} = 1 + (2 \times 3) = 7$. Will the actual observed value of y_{i+1} be exactly equal to 7? Probably not, this is because a prediction is a random variable and therefore will have an underlying probability distribution. We can however construct a $(1 - \alpha)\%$ confidence interval around the predicted value. Such an interval is known as a prediction interval. It can be interpreted along similar lines to a confidence interval, but will generally be much wider because of the inherent uncertainty of predicting the future. In the case of multiple regression the prediction is obtained by

$$\hat{y}_{i+1} = \hat{\alpha} + \hat{\beta}_1 x_{i+1}^1 + \hat{\beta}_2 x_{i+1}^2 + \cdots + \hat{\beta}_k x_{i+1}^k + \varepsilon_{i+1}.$$

9.4   BUILDING AND ESTIMATING MULTIPLE LINEAR REGRESSION MODELS IN *R*

The package MASS has a data frame Rubber which contains data from accelerated testing of tire rubber. The variables are loss (observations on the abrasion loss in gm/hr), hard (measured in 'Shore' units) and tens (which measures the tensile strength in kg/sq m). We are interested in abrasion loss given hardness and tensile strength. We examine the scatter plot matrix via the function plot(Rubber), illustrated in Figure 9.1. The scatter plot informs us that loss has a negative relationship with both hard and tens.
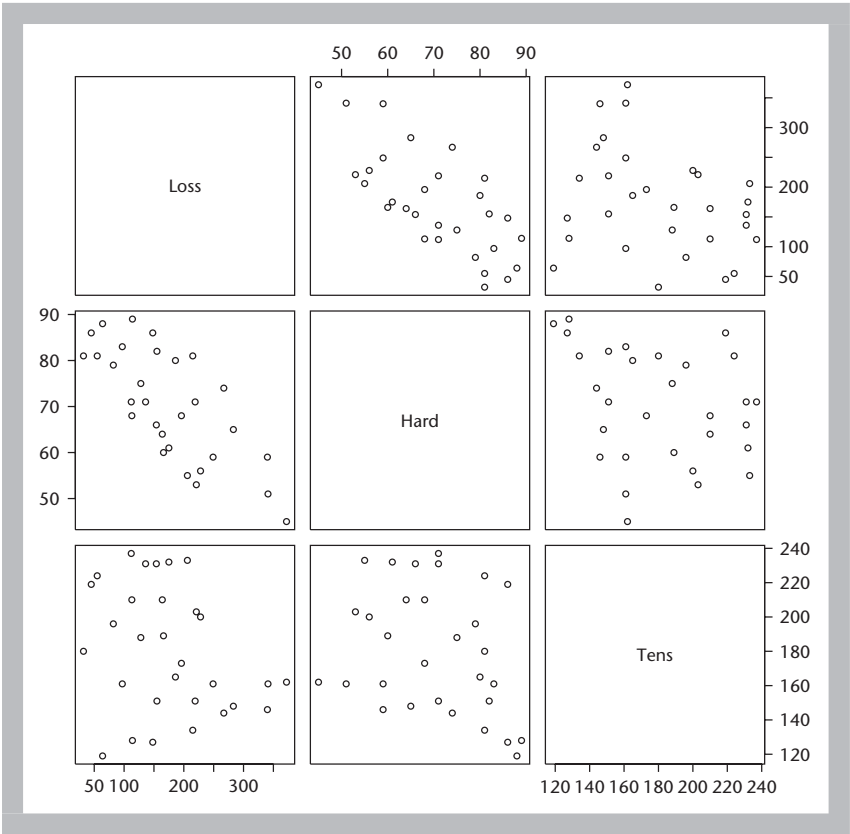
Figure 9.1 Scatter plot matrix of the variables in the data frame rubber

This is confirmed via the sample correlation coefficient

```
> cor(Rubber)
            loss          hard          tens
loss        1.000         -0.738        -0.298
hard        -0.738        1.000         -0.299
tens        -0.298        -0.299        1.000
```

We postulate the multiple regression model

$$loss = intercept - \beta_1 \, hard - \beta_2 \, tens + error$$

The model parameters can be estimated using `lm()`

```
> Rubber.reg1 <- summary(Rubber..reg <- lm(loss
~ hard+tens,data=Rubber))
```

```
> coef(Rubber.reg1)
            Estimate   Std.Error    t value   Pr(>|t|)
(Intercept)   885.16    61.752        14.33    3.84e-14
hard          -6.57      0.583       -11.27    1.03e-11
tens          -1.37      0.194        -7.07    1.32e-07
```

The coefficients have the expected sign, moreover the *t*-test for the intercept and both coefficients each reject the null hypothesis of equality to zero.

The previously mentioned statistics, alongside the parameter estimates, can be viewed using summary()

```
> summary(Rubber.reg)
Call:
lm(formula = loss ~ hard + tens, data = Rubber)
Residuals:
   Min     1Q Median     3Q     Max
-79.38 -14.61   3.82  19.75   65.98
Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  885.161   61.752       14.33  3.8e-14 ***
hard          -6.571    0.583      -11.27  1.0e-11 ***
tens          -1.374    0.194       -7.07  1.3e-07 ***
---
Signif. codes:   0 '***'      0.001   '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1

Residual standard error: 36.5 on 27 degrees of freedom
Multiple R-Squared: 0.84,    Adjusted R-squared: 0.828
F-statistic:   71 on 2 and 27 DF,  p-value: 1.77e-11
```

In this example the adjusted coefficient of determination is slightly lower than the standard coefficient of determination. Moreover, the *F*-test rejects the null hypothesis that all of the coefficients are zero and the coefficient *t*-tests are all significantly different from zero.

We can use the predict() function to forecast future values. For example using Rubber.reg as the argument we see that

```
> options(digits=4)
> Rubber.pred
<-predict(Rubber.reg,se.fit=TRUE,interval="prediction")
> Rubber.conf
<-predict(Rubber.reg,se.fit=TRUE,interval="confidence")
```

```
> Rubber.pred$fit[1,]
    fit          lwr            upr
    366.8        283.8          449.9
> Rubber.conf$fit[1,]
    fit          lwr            upr
    366.8        330.9          402.7
```

We print out the first prediction, alongside the upper and lower bounds, a prediction interval and confidence interval. The two returned intervals illustrate how if you use the argument `interval = ``prediction"` the function `prediction()` generates a prediction interval for the predicted value, while if you use the argument `interval = ``confidence"` a confidence interval around the prediction is returned.

In most cases we would not use all of the data to estimate the model, rather we might estimate the regression parameters using 70% of the available data, and investigate the model's predictive ability using the remaining 30% of data. It is possible to specify both an estimation and evaluation data set by including the arguments `data = estimation_data` and `newdata = testdata` in `predict()`

9.5   MULTIVARIATE REGRESSION

On occasion we may have q dependent variables $\{Y^1, Y^2, ..., Y^q\}$ which depend on a common set of independent variables $\{X^1, X^2, ..., X^k\}$. In this circumstance we wish to estimate the following regression equations

$$\begin{pmatrix} y_i^1 = \alpha_1 + \beta_{11}x_i^1 + \beta_{12}x_i^2 + \cdots + \beta_{1k}x_i^k + \varepsilon_{1i} \\ y_i^2 = \alpha_2 + \beta_{21}x_i^1 + \beta_{22}x_i^2 + \cdots + \beta_{2k}x_i^k + \varepsilon_{2i} \\ \vdots \\ y_i^q = \alpha_q + \beta_{q1}x_i^1 + \beta_{q2}x_i^2 + \cdots + \beta_{qk}x_i^k + \varepsilon_{qi} \end{pmatrix}.$$

If we assume the above regression equations are independent of each other, we can simply estimate a series of q regressions using `lm()` combined with `cbind()`.

The data frame `unemployment`[1] in the package `lmtest` contains data on the economic variables unemployment rate (`UN`), broad money supply (`m`), implicit deflator of Gross National Product (`p`), real purchases of goods and services (`G`) and real exports (`x`). We wish to estimate three separate
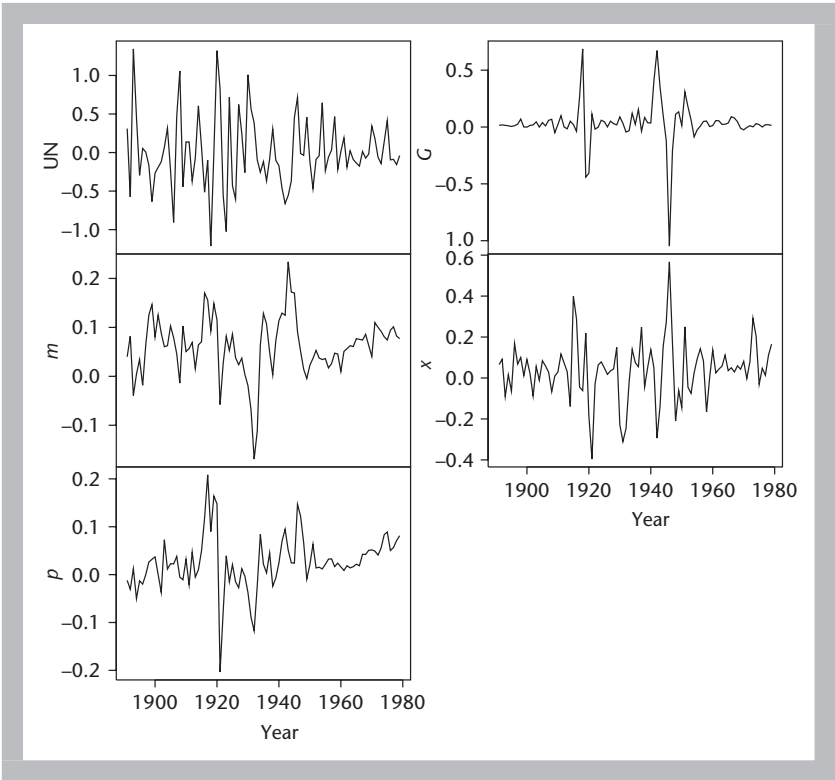
Figure 9.2 Time series plots of the log difference in macroeconomic variables from the data frame unemployment

linear regression models using the same independent variables

$$\begin{pmatrix} \Delta\text{UN} = \alpha_1 + \beta_{11}\Delta p + \Delta m + \varepsilon_{1i} \\ \Delta G = \alpha_2 + \beta_{21}\Delta p + \Delta m + \varepsilon_{2i} \\ \Delta x = \alpha_2 + \beta_{31}\Delta p + \Delta m + \varepsilon_{3i} \end{pmatrix}.$$

The symbol Δ indicates the log difference of the variable. The time series plot of each of the variable can be observed in Figure 9.2. We estimate the parameters for the regression equation as follows

```
>unemployment.change.reg=lm(cbind(UN,G,x)~p+m,data
 =unemployment.change)
> coef(unemployment.change.reg)
                 UN        G         x
(Intercept    0.2060    0.01341   -0.006041
p             0.8596   -0.85520    0.891759
m            -3.6518    0.65930    0.346579
```

We can also look in more detail at each individual regression

```
> summary(unemployment.change.reg)
Response UN:
Call:
lm(formula = UN ~ p + m, data = unemployment.change)
Residuals:
     Min      1Q   Median      3Q      Max
-1.15923 -0.23082 -0.00103  0.16956  1.40978
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.206      0.065    3.17  0.00213 **
p              0.860      0.999    0.86  0.39193
m             -3.652      0.945   -3.86  0.00022 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
Residual standard error: 0.423 on 86 degrees of
freedom
Multiple R-Squared: 0.176,  Adjusted R-squared: 0.157
F-statistic: 9.17 on 2 and 86 DF,  p-value: 0.000246
Response G:
Call:
lm(formula = G ~ p + m, data = unemployment.change)
Residuals:
     Min      1Q   Median      3Q      Max
-0.99546 -0.04235 -0.00654  0.03451  0.68991
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0134     0.0282    0.48    0.636
p            -0.8552     0.4336   -1.97    0.052 .
m             0.6593     0.4103    1.61    0.112
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
Residual standard error: 0.184 on 86 degrees of
freedom
Multiple R-Squared: 0.0462, Adjusted R-squared: 0.024
F-statistic: 2.08 on 2 and 86 DF,  p-value: 0.131
Response x:
Call:
lm(formula = x ~ p + m, data = unemployment.change)
Residuals:
    Min      1Q  Median      3Q      Max
-0.4153 -0.0514  0.0135  0.0569  0.4105
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.00604    0.02019   -0.30   0.7655
p            0.89176    0.31005    2.88   0.0051 **
m            0.34658    0.29338    1.18   0.2407
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
Residual standard error: 0.131 on 86 degrees of
freedom
Multiple R-Squared: 0.201,  Adjusted R-squared: 0.183
F-statistic: 10.8 on 2 and 86 DF,  p-value: 6.37e-05
```

We now see that cbind() forces lm() to estimate separate regression equations using the same independent variables. If the regression equations are not independent then in theory the model parameters should be estimated simultaneously because, provided the model is correctly specified, the estimators will have smaller standard errors than those obtained by estimating each regression equation separately. Estimation requires use of multivariate estimation techniques such as univariate least absolute deviation or coordinate rank regression. Fortunately estimation methods for multivariate regression can be found in the R package pls.pcr which can be freely downloaded from http://www.r-project.org/

9.6   SUMMARY

Multiple regression is an extension of simple regression where there are more than one independent variable. It is used in circumstances where it is believed that a dependent variable is influenced by more than one independent factor. The model can be assessed using t-tests on the individual coefficients, an F-test that all coefficients are significant and/or via the adjusted R^2 statistic. Estimation of the model parameters can be easily carried out in most statistical programs and spreadsheet packages.

9.7   FURTHER READING

Multiple regression is discussed in Doran (1989), Weisberg (1985), and Neter *et al*. (1996). Applications in risk management alongside Excel spreadsheets and Visual Basic source code are discussed in Lewis (2004).

Doran, H. E. (1989) *Applied Regression Analysis in Econometrics*, Marcel Dekker, Inc., New York.

Lewis, Nigel Da Costa (2004) *Operational Risk with Excel and VBA: Applied Statistical Methods for Risk Management*, John Wiley & Sons, Inc., New York.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996) *Applied Linear Regression Models* (3rd edn), Richard D. Irwin, Inc., Chicago, IL.

Weisberg, S. (1985) *Applied Linear Regression*, John Wiley and Sons, New York.

9.8   REVIEW QUESTIONS

1 Explain the difference between simple linear regression and multiple linear regression.

2 How might you assess the validity of a multiple regression model?

3 What is the relationship between the product moment correlation coefficient and R-squared?

4 Give the form of the following regression models:
   (a) Exponential regression
   (b) Logarithmic regression
   (c) Logistic regression

Misspecification Testing

Arguably, one of the most important issues in regression modeling applied to risk management is the correct specification of the regression equation. How can we assess the validity of the pre-specified regression model, which will provide the basis of statistical inference and practical decisions? It turns out that statistical inference concerning the linear regression model depends crucially on the "validity" of the underlying statistical assumptions of the model. If the assumed underlying statistical assumptions are invalid the inference based on it will be unreliable. The primary objective of this chapter is to outline the key assumptions of the linear regression model and provide some elementary techniques for validating or refuting these assumptions given a specific data set.

10.1 ASSUMPTIONS OF LINEAR REGRESSION

The statistical validity of linear regression analysis rests on the following assumptions:

- *Linearity*: A linear relationship exits between the dependent and independent variable. If this assumption is not met the model will not be adequate and the OLS estimator will be biased.

- *Homoscedasticity*: The variance of the sample is constant and does not change as the independent variable increases or decreases. This implies the variance of the residual term σ^2 is constant across all values of X. This is known as homoscedasticity in variance. A violation of this assumption is called heteroscedasticity.

- *Normality*: The residual is normally distributed.

■ *Independent variables uncorrelated*: The independent variables should ideally be uncorrelated. A violation of this assumption causes the least squares estimator to be inefficient. Inefficiency implies estimates with highly correlated independent variables will have a larger standard error than what they will have if the variables were uncorrelated. High correlation between the independent variables is sometimes termed as multicollinearity.

■ *No autocorrelation*: The dependent variable and error terms are independent, identically distributed. This ensures that errors associated with different observations are independent of one another. It also implies the residuals are not correlated with the dependent variable. A violation of this assumption is called auto or serial correlation.

If all of the above assumptions are valid, then the OLS estimators are known as the Best Linear Unbiased Estimators (BLUE). A BLUE estimator has the smallest variance in the class estimators that are linear in the dependent variable. Why is this useful? Because apart from wanting an estimator to be unbiased, we would also like an estimator that is always close to the population parameter we are trying to estimate. One way to measure this closeness is through the standard error of the estimator. If we have two unbiased estimators one with a large standard error and the other with a smaller standard error we would always select the latter.

Once the regression model has been constructed we will need to investigate how well it satisfies the above assumptions. Significant violations imply the regression model is misspecified. In the following sections we discuss some simple methods for assessing whether or not the regression assumptions are satisfied. We also outline what to do if we find serious violation of the assumptions.

10.2 LINEARITY

The easiest way to check for linearity is via a scatter plot of the dependent against the independent variables. A linear relationship should be evident. Any observable non-linear pattern is indicative of a violation of the linearity assumption. We have two choices when linearity fails. The first is to use a non-linear regression model, the second is to transform the data in an attempt to make it linear. The most popular linearizing transformation is to take the natural logarithm or square root of the dependent and/or independent variables. Both of these transformations are part of the family of

power transformation

$$Y(\lambda) = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, \\ \log(Y), & \text{if } \lambda = 0. \end{cases}$$

The above transformations are often known as Box Cox transformations.

10.3   HOMOSCEDASTICITY

Plotting the standardized residuals is a quick and simple way to check for a violation of the homoscedasticity assumption. If the variance is homoscedastic we would expect the scatter of points to be distributed randomly around zero, with no discernable trend. If this is not the case, it is indicative of a violation of this assumption. An alternative is to use a test statistic such as the Breusch Pagan test.

10.4   NORMALITY

Linear regression and many other statistical tests are based on the assumption of normality. The assumption of normality often leads to procedures that are simple and mathematically tractable compared to procedures that do not make the normality assumption. A plot of the histogram of the standardized residuals should be approximately symmetric and bell-shaped. We could also calculate the mean and median, which should be equal, and the skew (which should be zero) and kurtosis (which should be equal to 3). In addition to these simple tests, we could also use a QQ plot or even conduct a formal hypothesis test. If we find that the residuals violate the normality assumption, an appropriate transformation of the dependent variable can often yield residuals which are more approximately normally distributed. The Box Cox power transformation are frequently used to induce normality.

10.5   INDEPENDENT VARIABLES UNCORRELATED

This can be accomplished by calculating the correlation between the k independent variables. High correlation is indicative of a violation of this assumption. High correlation, say, between X^1 and X^2, implies that both variables are measuring similar aspects of the dependent variable and imply one can be substituted for the other. In other words we should re-specify the regression model with only X^1 (or X^2) but not both.

10.6   AUTOCORRELATION

The regression residuals are assumed to be uncorrelated with one another. We can investigate this by plotting the standardized residuals and observing whether or not a systematic pattern, indicative of autocorrelation, is evident. The consequences of proceeding with the OLS estimator may be very serious indeed. First, the OLS estimates are likely to be estimated with wide standard errors, which in turn imply that any hypothesis tests or inferential statements based on confidence intervals may be very misleading. Second, forecasts of future values of the dependent variable will have unnecessarily wide prediction intervals. Given the serious consequences of autocorrelation it is also advisable to employ formal statistical testing. The test most often used is the Durbin Watson test statistic

$$DW = \frac{\sum_{i=2}^{N} (\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{N} \varepsilon_i}.$$

The sampling distribution of this test statistic is somewhat complicated, depending on the particular values of the independent variables. Fortunately, it is known that values of dw lie between 0 and 4, and in general the closer the test statistic is to 2 the less likely a violation of the autocorrelation assumption. As a general rule of thumb, for samples of 100 observations or more, we can take high (or low) values of this test statistic as an indication that there is something wrong with the regression model. If we reject the null hypothesis of no autocorrelation the issue becomes what we should do. There are two possible approaches to resolve this issue. The first is to maintain that the regression model is correct and the observed autocorrelation happens to be a real property of the residuals. If this is indeed the case, an alternative method for estimation that allows for the observed pattern of autocorrelation has to be sought. The second approach is to acknowledge that the failure to reject the null hypothesis of no autocorrelation is a consequence of a misspecified model. If this is deemed to be the case, the solution is to re-specify the regression model. This will involve the addition of more independent variables and/or further transformations of the dependent and independent variables.

10.7   MISSPECIFICATION TESTING USING *R*

To illustrate some of the issues involved in misspecification testing we return to the simple regression model using data frame `cars` discussed in Chapter 8. We have already seen that the relationship between the dependent variable, `dist`, and the independent variable, `speed`, appear approximately linear. Next we investigate the issue of homoscedasticity.
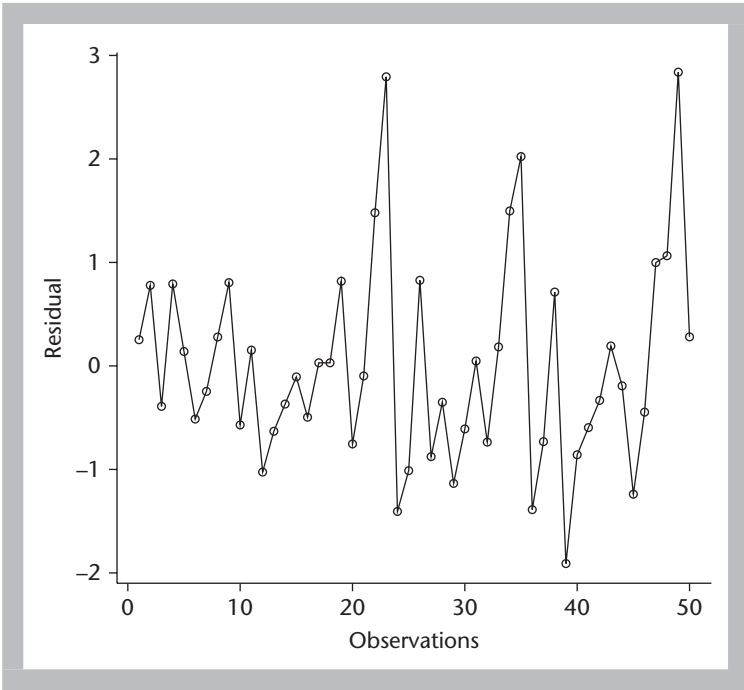
Figure 10.1 Standardized residuals from the regression of distances taken to stop on the speed of cars

We begin by plotting the standardized residuals, shown in Figure 10.1. It is not immediately obvious that a serious violation has occurred. However, close inspection reveals a cause for concern that as you move from observation 1 to 50 the variability in the observations seems to be increasing. Since the figure in itself is inconclusive we conduct the Breusch Pagan test for heteroscedasticity[1]

```
> lines(resid(car.reg))
> library(lmtest)
> bptest(dist~speed,data=cars)
          studentized Breusch-Pagan test
data: dist ~ speed
BP = 3.2149, df = 1, p-value = 0.07297
```

Given the result of the test and the evidence presented in Figure 13.4 we do not reject the assumption of homoscedasticity.[2]

Next we address the issue of normality. We begin by plotting the standardized residuals via qqnorm(), illustrated in Figure 10.2. The spread at the extreme percentiles is indicative of a fat tailed distribution. Our concern is
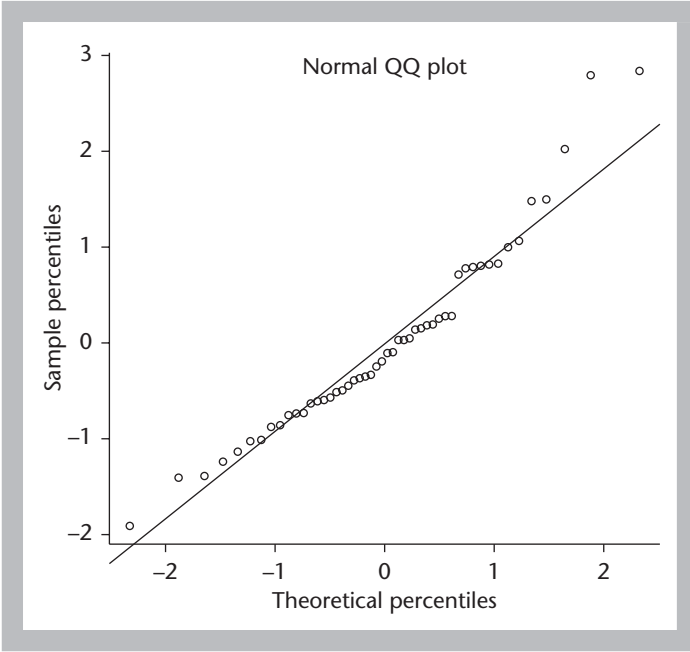
Figure 10.2 QQ normal plot of the standardized residuals from
the regression of distances taken to stop on the speed of cars

further raised when we look at the descriptive statistics via the user defined
function `Descriptive()`

```
> Descriptive(standard.resid)
 mean    med   range   IQR    sd    skew   kurt   max
0.000 -0.149  4.748 1.231 1.000 0.859  3.739  2.838
   min
-1.910
```

In this case the mean lies away from the median, indicating a degree of mild
skew. More important is the degree of kurtosis which at 3.739 is indicative
of non-normality. Given these indications of non-normality, we conduct a
formal hypothesis test using `shapiro.test()`

```
> shapiro.test(standard.resid)
      Shapiro-Wilk normality test
data: standard.resid
W = 0.9451, p-value = 0.02153
```

The *p*-value confirms our suspicions, and we reject the null hypothesis of
normality at the 5% significance level. In order to correct this problem[3] we

take the natural logarithm of both the dependent and independent variable

```
> car.reg1< −lm(log(dist)~log(speed),data=cars)
>standard.resid1=(resid(car.reg1)-
mean(resid(car.reg1)))/sd(resid(car.reg1))
```

We then take a look at the descriptive statistics

```
> Descriptive(standard.resid1)
  mean     med range    IQR    sd   skew  kurt   max    min
 0.000  -0.072 4.700 1.129 1.000 -0.032 2.867 2.201 -2.498
```

The results appear positive, very little skew, and a kurtosis close to 3. We now carry out the Shapiro–Wilk test

```
> shapiro.test(resid(car.reg1))
  Shapiro-Wilk normality test
data: resid(car.reg1)
W = 0.9911, p-value = 0.9684
```

This time we do not reject the null hypothesis. The correct specification, requires that we estimate the regression using the natural logarithm of both variables.

Finally we use `dw.test()` to investigate autocorrelation in the new specification

```
> dwtest(log(dist)~log(speed),data=cars)
         Durbin-Watson test
data: log(dist) ~ log(speed)
DW = 2.4115, p-value = 0.9088
```

Fortunately the specification appears not to reject the null hypothesis.

10.8 SUMMARY

Given the widespread popularity and use of regression models, it is of paramount importance to have a good insight into the tools available for testing the underlying statistical assumptions. The field of misspecification testing concentrates on appropriate diagnostic checking of the regression models. This chapter has offered an elementary review of the most import-ant assumptions and outlined simple techniques that can be used to assess them. The key areas to test are linearity, homoscedasticity, and normality.

It is also necessary to see that the independent variables are uncorrelated and that the model does not exhibit autocorrelation.

10.9 FURTHER READING

Further discussion about the assumptions of the linear regression model alongside an extensive battery of statistical tests can be found in Doran (1989), Weisberg (1985), and Neter *et al.* (1996).

Doran, H. E. (1989) *Applied Regression Analysis in Econometrics*, Marcel Dekker, Inc., New York.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996) *Applied Linear Regression Models* (3rd edn), Richard D. Irwin, Inc., Chicago, IL.

Weisberg, S. (1985) *Applied Linear Regression*, John Wiley and Sons., New York.

10.10 REVIEW QUESTIONS

1 Explain the difference between simple linear regression and multiple linear regression in terms of the underlying statistical assumptions.

2 Suppose you estimate a multiple regression model and find the $R^2 = 99.99\%$, but the model violates the homoscedasticity and normality assumption. Explain what (if anything) you should do next.

3 Using the data frame `cars` estimate via OLS the model parameters and then report and interpret an appropriate test statistic for each of the following:
   (1) Homoscedasticity
   (2) Homogeneity
   (3) Linearity
   (4) Normality

Non-linear and Limited Dependent Regression

There is no guarantee that the relationship between the dependent and independent variables will be linear. On many occasions we may find the relationship to have considerable non-linearity. In such circumstances, we might attempt to use polynomial regression, logarithmic regression, exponential regression, or a more general non-linear model. This chapter introduces these models. It also discusses the use of limited dependent regression models.

11.1 POLYNOMIAL REGRESSION

In the polynomial regression model we begin by postulating that the relationship between the dependent and independent variables can be modeled using a quadratic regression model of the form

$$y_i = \alpha + \beta_1 x_i + \beta_2 (x_i)^2 + \varepsilon_i.$$

If after suitable diagnostic checks the quadratic regression model proved unsatisfactory we would then consider a model which includes the independent variable X cubed

$$y_i = \alpha + \beta_1 x_i + \beta_2 (x_i)^2 + \beta_3 (x_i)^3 + \varepsilon_i.$$

If this model proved inadequate we would try

$$y_i = \alpha + \beta_1 x_i + \beta_2 (x_i)^2 + \beta_3 (x_i)^3 + \beta_4 (x_i)^4 + \varepsilon_i,$$

and so on. Polynomial regression analysis is sequential. We first evaluate a quadratic model, if this is inadequate we add a cubic term, and then decide whether or not the addition of such a term is justified.

11.2   LOGARITHMIC, EXPONENTIAL, AND MORE GENERAL FORMS OF NON-LINEAR REGRESSION

Logarithmic regression takes the dependent variable to be a function of the natural logarithm of the independent variables

$$y_i = \alpha + \beta_1 \ln(x_i) + \varepsilon_i.$$

Exponential regression takes the dependent variable to be a function of the exponential of the

$$y_i = \alpha \exp(\beta_1 x_i) + \varepsilon_i.$$

A more general form of non-linear regression takes the form

$$y_i = f(\beta_1 x_i) + \varepsilon_i,$$

where $f(.)$ is some postulated function specified by the risk analyst.

11.3   NON-LINEAR REGRESSION MODELING USING *R*

Non-linear least squares, implemented in the function `nls()`, can often be used to successfully estimate non-linear models. We illustrate this by returning the *R* data frame `cars` discussed in Chapter 8. A quadratic regression model for this data would be given by

```
> car.reg.cubic<-lm(dist~speed+I(speed^2),data=cars)
> summary(car.reg.cubic)
Call:
lm(formula = dist ~ speed + I(speed^2), data = cars)
Residuals:
   Min      1Q Median      3Q     Max
-28.72   -9.18   -3.19    4.63   45.15
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.470     14.817    0.17     0.87
speed           0.913      2.034    0.45     0.66
I(speed^2)      0.100      0.066    1.52     0.14
```

```
Residual standard error: 15.2 on 47 degrees of
freedom
Multiple R-Squared: 0.667,  Adjusted R-squared: 0.653
F-statistic: 47.1 on 2 and 47 DF,  p-value: 5.85e-12
```

The model does not appear to be adequate; perhaps a more general non-linear specification may improve things

```
> car.reg.nls<-nls(dist~a+speed^b,start=list(a=5,b=5),data=cars)
> summary(car.reg.nls)
Formula: dist ~ a + speed^b
Parameters:
  Estimate Std. Error t value Pr(>|t|)
a  -3.3568    4.4496   -0.75     0.45
b   1.3911    0.0297   46.83   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 15.1 on 48 degrees of freedom
Correlation of Parameter Estimates:
       a
b -0.877
```

The function nls() provides estimates of the parameters alongside their correlation. Note that there is no guarantee that the iterative procedure of this function will converge. It is therefore always useful to re-run the function with different starting values

```
> car.reg.nls<-
nls(dist~a+speed^b,start=list(a=50,b=15),data=cars)
> summary(car.reg.nls)
Formula: dist ~ a + speed^b
Parameters:
  Estimate Std. Error t value Pr(>|t|)
a  -3.3570    4.4496   -0.75     0.45
b   1.3911    0.0297   46.83   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
Residual standard error: 15.1 on 48 degrees of
freedom
```

```
Correlation of Parameter Estimates
       a
b -0.877
```

Fortunately, in this case the estimates converge.

11.4   LOGISTIC AND OTHER LIMITED DEPENDENT REGRESSION MODELS

Logistic regression is often used for analyzing data where the dependent variable is a binary response such as presence or absence of a disease or the success or failure of students in a university examination. Given a binary dependent variable Y, where $Y = 1$, with probability p, if an event of interest occurs, 0 otherwise, the observed number of successes are converted into proportions which are then fitted by a logistic model of the form

$$\log\left[\frac{p}{(1-p)}\right] = \alpha + \beta_1 x_i^1 + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \varepsilon_i.$$

This equation is very similar to the linear regression as $\{X^1, X^2, ..., X^k\}$ are the k independent variables and ε is the residual. However, unlike linear regression we do not require assumptions about normality of the residual term. In addition, using the logistic model, we are able to determine the probability success from the ratio, known as the odds ratio

$$\left[\frac{p}{(1-p)}\right].$$

The natural logarithm of the odds ratio is called the log odds ratio or logit. To see the value of this ratio notice the consequence of an odds ratio equal to 1

$$\left[\frac{p}{(1-p)}\right] = 1 \quad \Rightarrow \quad p = (1-p) \quad \Rightarrow \quad p + p = 1 \quad \Rightarrow \quad p = 0.5.$$

Thus an odds ratio equal to 1 implies the probability that $Y = 1$ is 0.5. We interpret $\exp(\beta_i)$ as the effect of the independent variable on the odds ratio. For example, if we postulate the logistic regression

$$\log\left[\frac{p}{(1-p)}\right] = \alpha + \beta_1 x_i + \varepsilon_i,$$

and on estimation find that $\hat{\beta}_1 = 0.963$ then $\exp(\hat{\beta}_1) = 1.999$ then a 1 unit change in X would make the event Y about twice as likely.

The following R code illustrates the relationship between p, the odds ratio and the log odds ratio

```
> p=seq(from =0,to=1, by=.1)
> odd.ratio= p/(1-p)
> odds.ratio=p/(1-p)
> log.odds=log(odds.ratio)
> table=cbind(p,odds.ratio,log.odds)
> table
           p         odds.ratio     log.odds
  [1,]  0.0          0.0000         -Inf
  [2,]  0.1          0.1111         -2.1972
  [3,]  0.2          0.2500         -1.3863
  [4,]  0.3          0.4286         -0.8473
  [5,]  0.4          0.6667         -0.4055
  [6,]  0.5          1.0000          0.0000
  [7,]  0.6          1.5000          0.4055
  [8,]  0.7          2.3333          0.8473
  [9,]  0.8          4.0000          1.3863
 [10,]  0.9          9.0000          2.1972
 [11,]  1.0            Inf             Inf
```

We see that if $p > 0.5$ the odds ratio is greater than 1 and the log odds ratio is greater than zero. For $p < 0.5$ the odds ratio is less than 1 and the log odds ratio is negative. Since p is a probability the logistic regression model is constructed so that $0 \le p \le 1$. To see that this is so note that as

$$\alpha + \beta_1 x_i^1 + \beta_2 x_i^2 + \cdots + \beta_k x_i^k$$

becomes very large, p approaches 1 and as

$$\alpha + \beta_1 x_i^1 + \beta_2 x_i^2 + \cdots + \beta_k x_i^k$$

becomes very small, p approaches 0. Furthermore if

$$\alpha + \beta_1 x_i^1 + \beta_2 x_i^2 + \cdots + \beta_k x_i^k = 0$$

then $p = 0.5$.

11.4.1   Logistic regression model using *R*

Logistic regression coefficients can be estimated via the `glm()` function. We illustrate this using simulated data

```
> X <-rnorm(100)
> Y<-X+rnorm(100)+1
> Y <- as.integer(Y<0)
```

The above code, simulates 100 observations from a standard normal distribution and stores the results in X. The variable Y is generated in a similar fashion but is also linked to X. It is then converted into a binary variable using `as.integer()`. We can now estimate the logistic regression parameters

```
> Logistic.reg<-glm(Y~X,family=binomial(link=logit))
> summary(Logistic.reg)
Call:
glm(formula = Y ~ X, family = binomial(link = logit))
Deviance Residuals:
   Min     1Q  Median     3Q     Max
-1.411  -0.702  -0.426  -0.188   2.303
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.586      0.313   -5.07  4.1e-07 ***
X             -1.431      0.391   -3.66  0.00025 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 105.38  on 99  degrees of freedom
Residual deviance:  86.48  on 98  degrees of freedom
AIC: 90.48
Number of Fisher Scoring iterations: 4
```

In actual fact the `glm()` function can be used to estimate a number of regression models, the key argument to be aware of is `family\(=\)binomial(link\(=\)logit)`. For example, to estimate linear regression model we would type something like

```
> Linear.reg<-glm(Y~X,family=gaussian(link=identity))
```

11.4.2   Probit regression

An often used alternative to Logistic regression is probit regression

$$\Theta(p)^{-1} = \alpha + \beta_1 x_i^1 + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \varepsilon_i,$$

where $\Theta(p)^{-1}$ is the inverse cumulative normal probability function. In actual fact the shapes of the probit and logistic functions are very similar, except in the tails. We can estimate the probit model by using the argument `family=binomial(link=probit)` in `glm()`. For the data discussed previously, the parameters are estimated as

```
> Probit.reg<-glm(Y~X,family=binomial(link=probit))
> summary(Probit.reg)
Call:
glm(formula = Y ~ X, family = binomial(link = probit))
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.389  -0.718  -0.421  -0.138   2.329
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.928      0.168   -5.52  3.5e-08 ***
X             -0.830      0.214   -3.88  0.00010 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 105.382  on 99  degrees of freedom
Residual deviance:  86.279  on 98  degrees of freedom
AIC: 90.28
Number of Fisher Scoring iterations: 4
```

11.4.3   Complementary log log regression

Complementary log log regression is occasionally used in place of logistic or probit regression. The model takes the form

$$\log(-\log(1-p)) = \alpha + \beta_1 x_i^1 + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \varepsilon_i.$$

In actual fact the model is similar to the logistic and probit models for values of p near 0.5, but differs substantially from them for values near 0 or 1. We can estimate the parameters using the argument `family=binomial(link=probit)` in `glm()`. For the data discussed

previously, the parameters are estimated as

```
> Cloglog.reg <- glm(Y ~ X,family=binomial(link= cloglog))
> summary(Cloglog.reg)
Call:
glm(formula = Y ~ X, family = binomial(link = cloglog))
Deviance Residuals:
   Min     1Q Median     3Q    Max
-1.477  -0.685  -0.434  -0.213   2.277
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.712      0.280   -6.11  9.8e-10 ***
X             -1.214      0.298   -4.07  4.6e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 105.38  on 99  degrees of freedom
Residual deviance:  86.23  on 98  degrees of freedom
AIC: 90.23
Number of Fisher Scoring iterations: 4
```

11.4.4   Other regression models in *R*

The function `glm()` can be used to estimate a large number of generalized linear models. The full list is given below

- ```gaussian(link = "xxx")``` accepts the links ```"identity", "log"``` and ```"inverse"```;

- ```binomial(link = "xxx")``` accepts the links ```"logit", "probit", "log"``` and ```"cloglog"```

- ```Gamma(link = "xxx")``` accepts the links ```"inverse", "identity"``` and ```"log"```

- ```inverse.gaussian(link = "xxx")``` accepts the links ```"log", "identity"```, and ```"sqrt"```

- ```inverse.gaussian(link = "1/mu^2")``` accepts the links ```"1/mu^2", "inverse", "inverse"``` and ```"log"```.

The common theme in all of these models is the analysis of relationships between multiple measurements made on groups of subjects or other objects. The relationship is analyzed by specifying a dependent variable and a number of independent variables.

11.5   SUMMARY

Statistical analysis is often focused on questions involving the relationship between one or more variables that are thought as being dependent on some other independent variables. There is not particular reason to suppose that the relationships will be linear. Non-linear regression models therefore offer an additional degree of flexibility. The statistical package *R* contains powerful algorithms, which make non-linear modeling as easy to use as linear regression modeling. Where the dependent variable is binary or only takes on a limited number of values, limited dependent regression models can be used. Again, these models can be easily estimated in *R*.

11.6   FURTHER READING

Further discussion of non-linear models can be found in Berndt *et al.* (1974), White (1981), Gallant (1987), Bates and Watts (1988), Cook and Weisberg (1990), Wolak (1991). Detailed discussion of Limited-dependent and regression modeling is given in Maddala (1983).

Bates, Douglas M. and Watts, Donald G. (1988) *Nonlinear Regression Analysis and Its Applications*, John Wiley & Sons, New York.

Berndt, E., Hall, B., Hall, R., and Hausman, J. (1974) "Estimation and inference in nonlinear structural models," *Annals of Economic and Social Measurement*, 3: 653–65.

Cook, R. D. and Weisberg, S. (1990) "Confidence curves in nonlinear regression," *Journal of the American Statistical Association*, 85: 544–51.

Gallant, A. R. (1987) *Nonlinear Statistical Models*, Wiley, New York.

Maddala, G. S. (1983) *Limited-dependent and qualitative variables in economics*, Cambridge University Press, Cambridge.

White, H. (1981) "Consequences and detection of misspecified nonlinear regression models." *Journal of the American Statistical Association*, 76: 419–33.

Wolak, Frank. (1991) "The local nature of hypothesis tests involving inequality constraints in nonlinear models," *Econometrica*, 59: 981–95.

11.7   REVIEW QUESTIONS

1 Give the form of the following regression models:
   (a) Exponential regression
   (b) Logarithmic regression
   (c) Logistic regression

2 Estimate a cubic regression model using the data frame `cars`. How does the model compare to the quadratic model estimated in this chapter?

3 Contrast logistic regression against probit regression. What are their similarities and differences?

Modeling Energy Price Volatility

Say the words "Energy price volatility" and many people will think of the OPEC oil price hikes of the 1970s or perhaps the more recent 2004 sharp upswing in the price of crude oil. Yet price volatility is a characteristic of capitalism and freely operating energy markets are no exception. Accurate estimates of the variation in energy asset values over time are important for the valuation of financial contracts, retail obligations, physical assets, and in solving portfolio allocation problems. As a consequence, modeling and forecasting of price volatility has acquired an unprecedented significance in the industry. In response, various attempts have been made to develop statistical tools to help characterize and predict price volatility. In general these models fall into three categories, Exponentially Weighted Moving Average models, Generalized Autoregressive Conditional Hetroscedasticity models, and Stochastic Volatility Differential Equations. In this chapter we introduce the first two of these modeling approaches.

12.1   THE CONSTANT VOLATILITY MODEL

Figure 12.1 shows time series plots of the price of the prompt (nearest) month West Texas Intermediate crude oil futures contract (top), 1-month natural gas forward (middle) and the spot price of unleaded gasoline (bottom) over the period February 1997 to March 2004. For all of the charts there are periods when the price rises or falls extremely quickly, and periods of relative stability. This is emphasized in Figure 12.2, which show the time series plots of the log returns for each of these products.
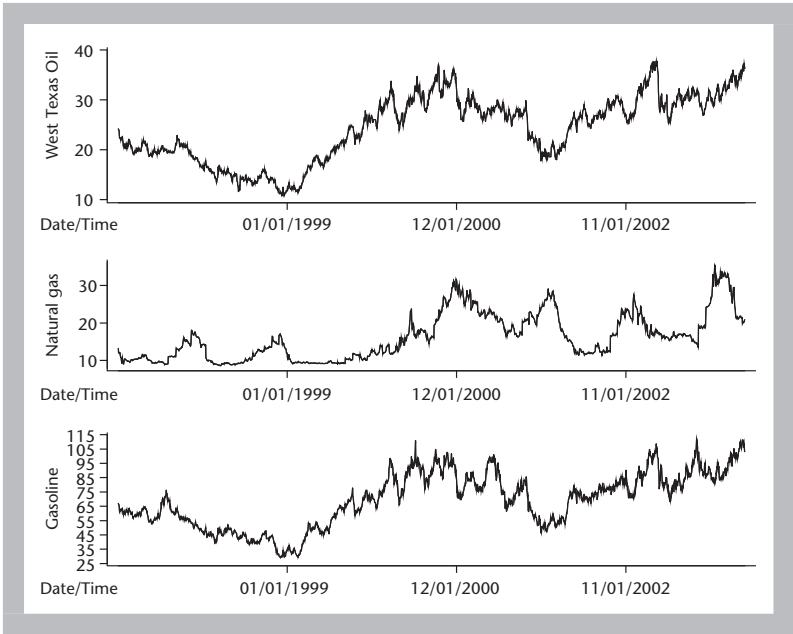
Figure 12.1 Time series plot of the prompt (nearest) month West Texas Intermediate crude oil contract (top), 1-month natural gas forward (middle), and the spot price of unleaded gasoline (bottom)
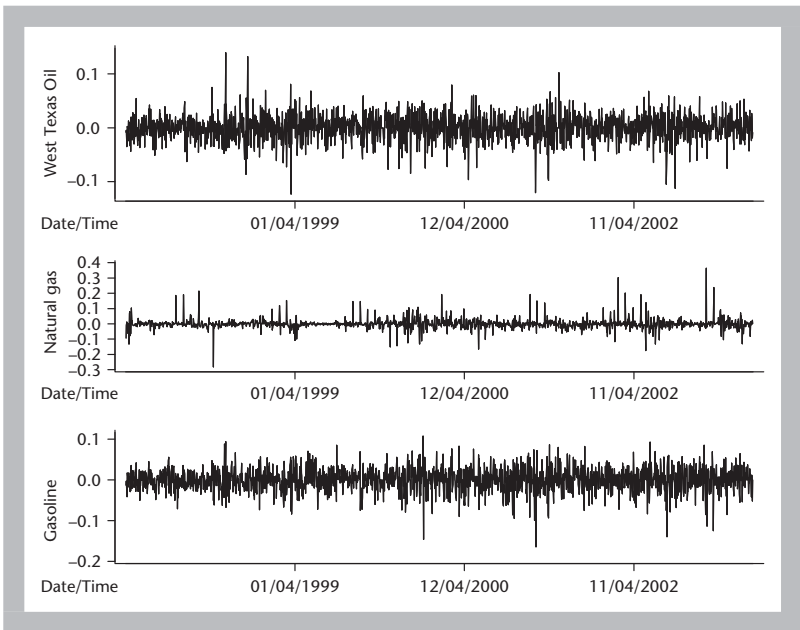
Figure 12.2 Time series plot of the log return of the prompt (nearest) month West Texas Intermediate crude oil contract (top), 1-month natural gas forward (middle), and the spot price of unleaded gasoline (bottom)
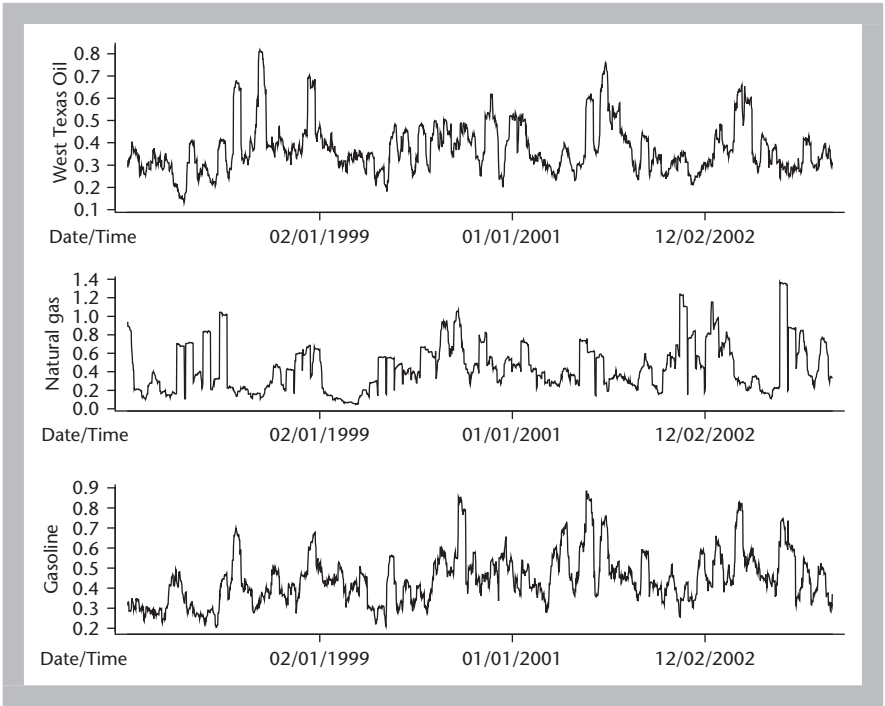
Figure 12.3 Annualized 20-day volatility time series plot of the prompt (nearest) month West Texas Intermediate crude oil contract (top), 1-month natural gas forward (middle), and the spot price of unleaded gasoline (bottom)

So far we have considered models for which the standard deviation (volatility) is constant. Casual inspection of Figures 12.1 and 12.2 does not appear to support this assertion. Indeed from our discussion of the economics of energy in Chapter 1 we know that the intensity of price movements in energy products is intimately linked to the ebb and flow of supply and demand. As such, a constant volatility model will be unduly simplistic for the vast majority of energy products. Figures 12.3 shows the annualized 20-day standard deviation for the three products. Again it appears to confirm our suspicion of periods of varying volatility. Figures 12.4, 12.5, and 12.6 show the annualized volatility of the prompt month West Texas Intermediate crude oil contract, 1-month natural gas forward and spot price of unleaded gasoline respectively. Imposed on each graph is the volatility trend line calculated using a locally weighted scatter-plot smoother. We can now see quite clearly that the trend in volatility is rising and falling over time with periods of high volatility clustering together and periods of low volatility clustering together.
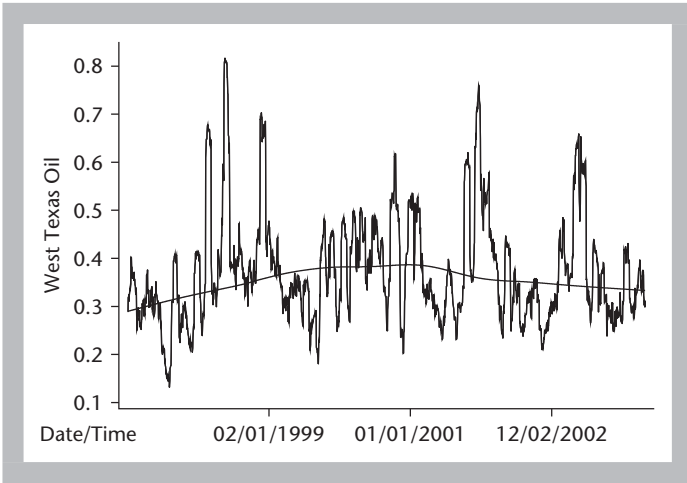
Figure 12.4 Annualized 20-day volatility time series plot of the prompt (nearest) month West Texas Intermediate crude oil contract with volatility trend line
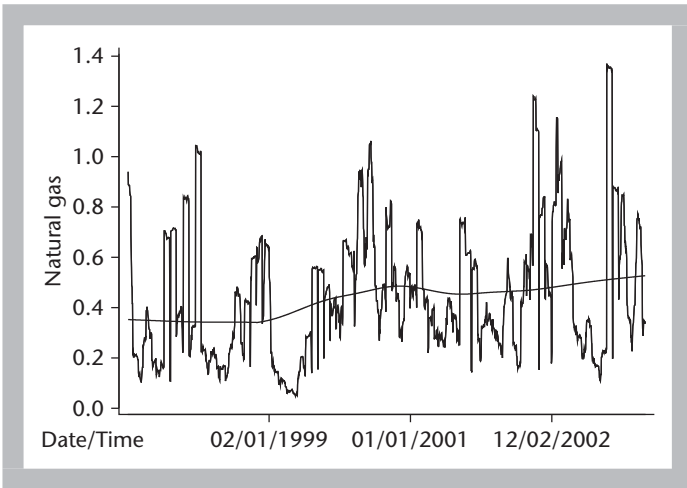
Figure 12.5 Annualized 20-day volatility time series plot of the 1-month natural gas forward with volatility trend line

We might wish to test the idea of constant volatility between two different time periods formally using statistical inference. A simple test statistic is computed as

$$\hat{T} = \frac{S_1^2}{S_2^2} \sim F_{N_1, N_2},$$
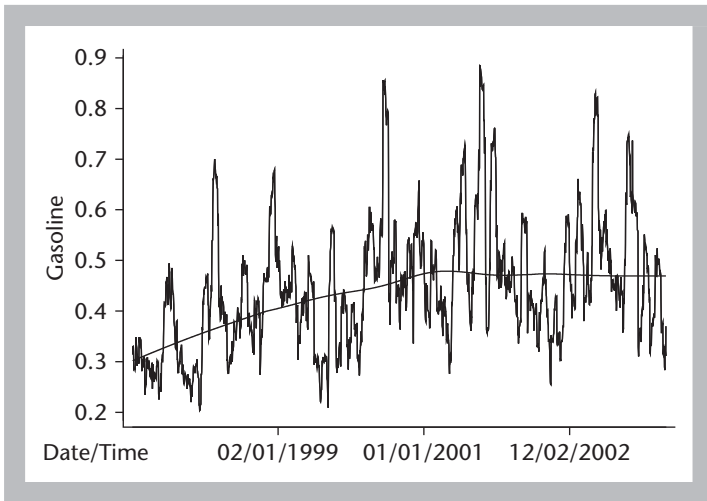
Figure 12.6 Annualized 20-day volatility time series plot of the spot price of unleaded gasoline with volatility trend line

where S_1^2 is the sample variance of the first time period and N_1 its corresponding sample size, S_2^2 the sample variance of the second period and N_2 its corresponding sample size and F_{N_1, N_2} the F distribution with N_1 and N_2 degrees of freedom. Critical values for the F distribution are given in Tables A.4–A.6.

Whilst a formal inferential statistical test has the reassuring aura of exactitude about it, we always need to beware that any such test may be powerless to detect problems of an unsuspected nature and may be heavily dependent upon assumptions about the actual distribution of volatility. Graphical techniques alongside formal statistical inference offer the best defense against unexpected structure.

12.2   EXPONENTIALLY WEIGHTED MOVING AVERAGE MODELS

Empirical data strongly indicates that the constant volatility model does not apply to energy products in particular, or financial asset prices in general. Indeed, there is considerable clustering of volatility at high and low levels as well as volatility persistence. Volatility persistence refers to the situation where high (low) volatility today tends to be followed by high (low) volatility tomorrow. Volatility clustering implies that volatility shocks today influence the expectation of volatility many periods into the future. Since accurate modeling volatility is critical for the implementation and evaluation of effective energy risk management, asset, and derivative pricing as

well as trading and hedging strategies, having access to the appropriate statistical tools is essential.

We saw in Chapter 3 that one measure of dispersion, the sample variance, is calculated by

$$S^2 = \frac{\sum_{i=1}^{N}(r_i - \bar{r})^2}{N - 1}, \text{ where } \bar{r} \text{ is the sample mean.}$$

We also observed for daily data that $\bar{r} \approx 0$. Therefore, for daily data, we could approximate sample volatility by

$$S^2 \approx \frac{\sum_{i=1}^{N} r_i^2}{N - 1}$$

Furthermore, since providing N is large, there is little difference between dividing by N or $N - 1$, we can write the approximation as

$$S^2 \approx \frac{\sum_{i=1}^{N} r_i^2}{N}$$

This formula provides the basis for a simple model of the volatility of energy price returns given by

$$\tilde{\sigma}_{t+1}^2 \approx \frac{\sum_{\tau=1}^{M} r_{t+1-\tau}^2}{M}$$

Thus the volatility forecast for day $t+1$ is calculated using the most recent M observations. Figure 12.7 shows this simple volatility estimator for 1-month natural gas forward with values of M ranging from 50 to 300. The general pattern of volatility observed is somewhat similar in all graphs, however there are significant differences in the forecast levels. This is particularly apparent if you compare the top graph ($M = 300$) with the bottom graph ($M = 50$). These differences raise the question of how we should determine M. The short answer is that there is no optimal value for M. Experimentation with different values is required. However, you should bear in mind that too large a value for M will make $\tilde{\sigma}^2$ unresponsive to current/recent price swings. Too small a value will make the volatility forecast too sensitive to current/recent price swings.

12.2.1   The standard EWMA estimator

An equally weighted estimator implies returns generated M periods ago are as important as a return generated 1 period ago. It can be argued that it is
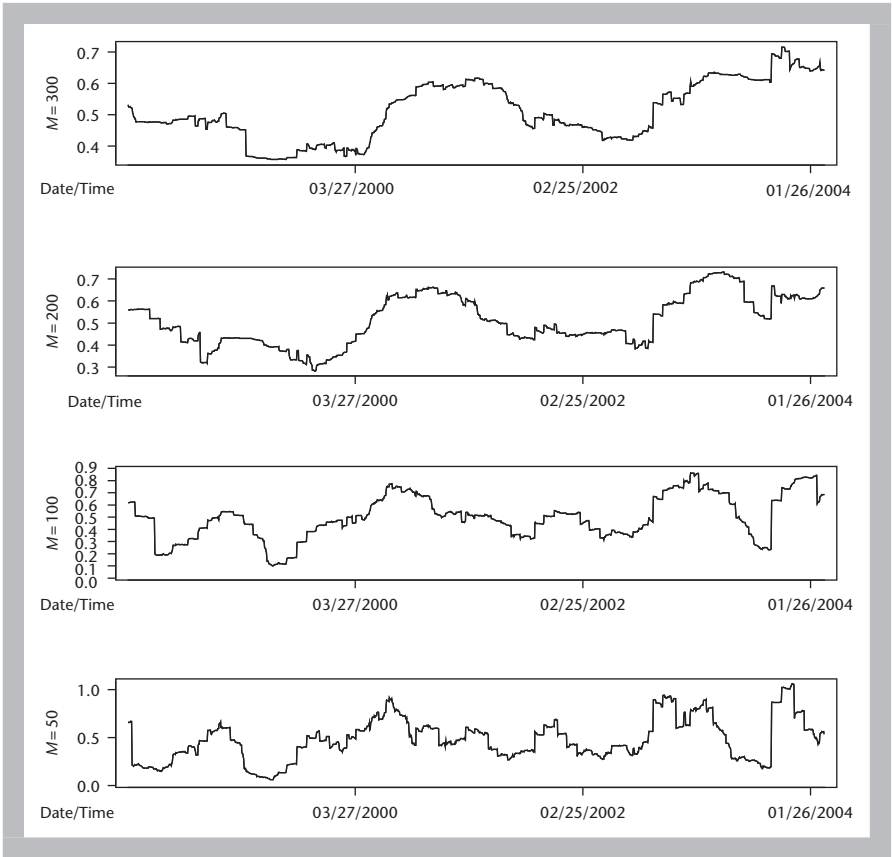
Figure 12.7 Simple volatility forecasts using various values for *M*

more reasonable to give a larger weight to recent observations than distant observations. This would better fit with our intuition that today's news is more important for energy prices than yesterday's news. Exponentially weighted moving average (EWMA) models capture volatility persistence in simple and flexible ways by giving more weight to recent observations than distant observations. EWMA estimates the conditional variance of a price return series $\{r_1, ..., r_N\}$ as

$$\tilde{\sigma}_{t+1}^2 = \lambda \tilde{\sigma}_t^2 + (1 - \lambda)r_t^2.$$

The EWMA estimator defines next period's volatility as a weighted average of this period's volatility ($\tilde{\sigma}_t^2$) and this period's squared return (r_t^2). In other words the volatility at $t + 1$ is a weighted average of the volatility at time t and the magnitude of the return at time t. The parameter λ is known as the decay factor. The smaller λ, the greater the weight given to recent returns.
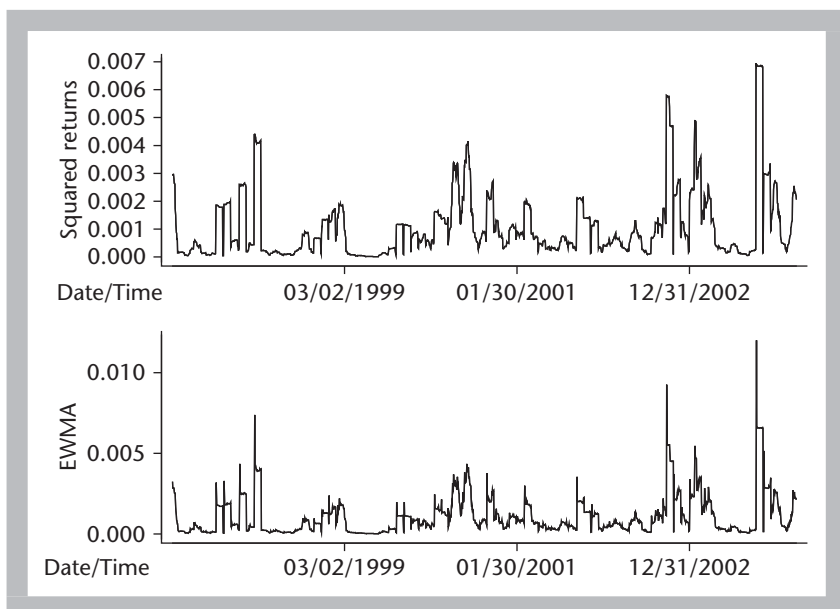
Figure 12.8 Squared returns of 1-month natural gas forward (top) and EWMA forecasts (bottom) with $\lambda = 0.96$

This can more easily be seen by writing the EWMA in an alternative form

$$\tilde{\sigma}_{t+1}^2 = (1 - \lambda) \sum_{i=1}^{\infty} \lambda^{i-1} r_{t+1-i}^2.$$

The decay factor can be estimated using the method of maximum likelihood or an alternative chosen on the basis of out of sample performance of the EWMA model against some predetermined criteria. Figure 12.8 shows the squared returns of the 1-month natural gas forward (top) alongside EWMA forecasts (bottom) with $\lambda = 0.96$.[1]

12.2.2   The robust–Laplace EWMA

An alternative exponentially weighted specification, known as a robust–Laplace exponentially weighted moving average (REWMA) uses an infinite exponentially weighted average of past absolute returns rather than the weighted average

$$\tilde{\sigma}_{t+1}^2 = \left\{ \lambda \sigma_t + (1 - \lambda)\sqrt{2}\, |r_t| \right\}^2.$$

This specification for volatility is derived using the maximum likelihood estimator for the variance of a Laplace distribution which has fatter tails and is considerably more peaked than the normal distribution. In fact the maximum likelihood estimator of the variance of the Laplace distribution is given by

$$\tilde{\sigma}^2 = \left\{ \frac{\sqrt{2}}{N} \sum_{i=1}^{N} |r_i| \right\}^2 .$$

It is therefore a function of the squared absolute return rather than the average square return.

Figure 12.9 gives the REWMA forecasts (top) and the EWMA forecasts (bottom) for the 1-month natural gas forward contract. Whilst there is some difference between the actual levels forecast, the general pattern is quite similar. In fact the correlation between the REWMA and EWMA forecasts is 0.78. One advantage of REWMA over EWMA is that it offers a more robust estimator. REWMA places less weight on observations which are either very large or very small than EWMA, with relatively more weight going to those observations with moderate values.

Both EWMA and REWMA estimators incorporate all of the past volatility shocks via the squared returns but with declining weights. For example, the
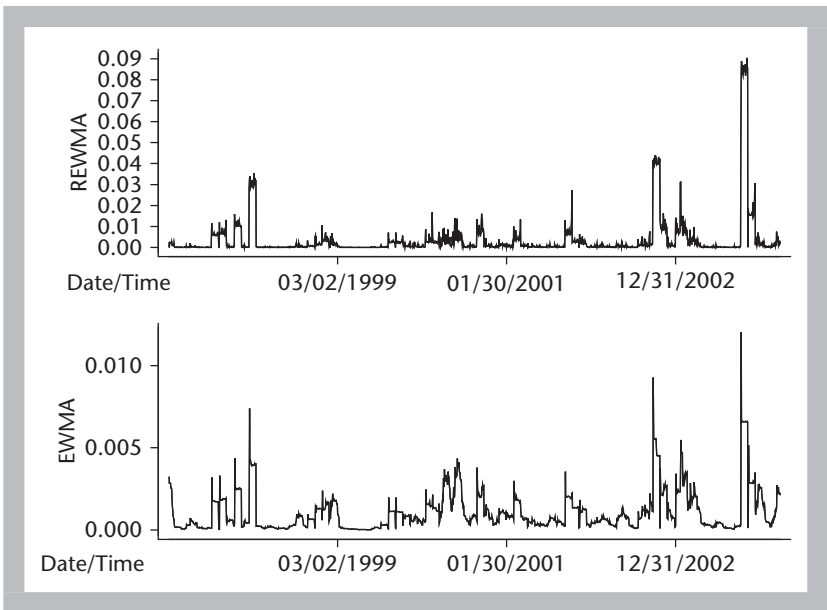
Figure 12.9 Comparison of REWMA (top) and EWMA (bottom) volatility forecasts of the 1-month natural gas forward contract with $\lambda = 0.96$

EWMA estimator can be written as

$$\tilde{\sigma}_{t+1}^2 = (1 - \lambda) \sum_{i=0}^{\infty} \lambda^i r_{t-i}^2,$$

and the REWMA estimator can be written as

$$\tilde{\sigma}_{t+1}^2 = \left\{ (1 - \lambda)\sqrt{2} \sum_{i=0}^{\infty} \lambda^i |r_{t-i}| \right\}^2.$$

Therefore, EWMA and REWMA estimators are based on an infinite series of weighted past squared returns. For most energy products this may seem an untenable assumption because it implies that events that occurred in the distant past are still relevant today. In actual fact this is not much of a disadvantage, provided we only use EWMA models for relatively short term forecasting. In practice EWMA/REWMA work particularly well especially when used to forecast volatility in the near future. At short distances into the future they track volatility changes in a way that is broadly consistent with the observed returns. For near future forecasts they frequently outperform the more complicated models to be discussed in Section 12.3. Furthermore, since there is only one parameter to estimate EWMA/REMA models are very attractive when measured against our core objective of parsimony.

12.3 GENERALIZED AUTOREGRESSIVE CONDITIONAL HETROSCEDASTICITY MODELS

One of the prominent tools for modeling volatility is the generalized autoregressive conditional hetroscedasticity (GARCH) model. The GARCH model captures volatility persistence in simple and flexible ways by assuming that future volatility is dependent on past levels of volatility. The GARCH(p,q) model takes the form

$$\tilde{\sigma}_{t+1}^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i r_{t+1-i}^2 + \sum_{j=1}^{p} \lambda_j \tilde{\sigma}_{t+1-j}^2$$

where p is the degree of generalized autoregressive conditional hetroscedasticity and q the degree of autoregressive conditional hetroscedasticity.[2] In theory the value of p and q can be determined by looking at the autocorrelation function and partial autocorrelation function of the square of residues. In practice GARCH(p,q) models are difficult to estimate and often provide little additional explanatory power when compared to the GARCH(1,1).

12.3.1 The GARCH(1,1) model

The GARCH(1,1) is widely used because it provides a good description of volatility for a large number of financial price series. It is defined as

$$\tilde{\sigma}_{t+1}^2 = \alpha_0 + \alpha r_t^2 + \lambda \tilde{\sigma}_t^2.$$

From the above equation we see that it dissects future volatility into three components. The first component is a constant represented by α_0. The second component is composed of the squared returns αr_t^2, and the third component the current level of volatility $\lambda \tilde{\sigma}_t^2$. Since volatility is positive we expect the parameters $\alpha_0, \alpha, \lambda$ to be greater than zero. The parameter λ is called the volatility persistence parameter; a high value today implies a high carryover effect of past to future volatility, while a low value implies a weak dependence on past volatility. At the extreme if $\lambda = 0$ then tomorrow's volatility is simply a function of the constant and today's squared return. If $\lambda < 1$ volatility depends on all past volatilities with geometrically declining weights. The sum of $\alpha + \lambda$ represents the influence on the forecast volatility of the squared return and current level of volatility. In practice $\alpha + \lambda$ is often close to 1. Values close to 1 are a sign of inertia in the effect of shocks on the volatility of returns.

We see therefore, that the parameters α and λ provide the mechanism by which GARCH(1,1) captures the empirically observed feature of volatility clustering. Provided $\alpha + \lambda < 1$ the unconditional variance or long term volatility of the model can be calculated as

$$\tilde{\sigma}^2 = \frac{\alpha_0}{1 - \alpha - \lambda}.$$

This provides a useful alternative representation of

$$\tilde{\sigma}_{t+1}^2 = \tilde{\sigma} + \alpha(r_t^2 - \tilde{\sigma}^2) + \lambda(\tilde{\sigma}_t^2 - \tilde{\sigma}^2).$$

This equation informs us that the GARCH(1,1) forecast volatility is a weighted average of the long term volatility, today's squared return, and today's volatility. It demonstrates quite clearly the way in which GARCH models capture the feature of volatility clustering. If the market is volatile today in the sense that volatility is above the long term trend, next period's forecast will be high. The actual level of volatility is intensified or offset in accordance with the magnitude of today's squared return. If, on the other hand, today's volatility is below the long term trend, forecast volatility will be low as well, unless today's squared return is particularly large. The above equation also demonstrates the notion of mean reversion in volatility; Provided $\alpha + \lambda < 1$, after a shock volatility will eventually return to its

long term mean. If $\alpha + \lambda = 1$ volatility does not exhibit mean reversion but volatility persistence.

If we impose the conditions that $\alpha_0 = 0$ and $\alpha + \lambda = 1$, where α is equal to 1 minus the value of the EWMA λ, then the GARCH(1,1) model is known as an integrated generalized conditional hetroscedasticity (IGARCH) model and is in effect the EWMA estimator of Section 12.2. We can see now that under EWMA the long term volatility is ill-defined. This underscores our previous discussion that EWMA should be used for short term forecasting only. Furthermore, in IGARCH/EWMA a shock in the current period persists indefinitely. This is not the case with the GARCH(1,1) or GARCH(p,q) where a shock in the current period has a transient effect on volatility and dies out rapidly at an exponential rate of decay.

12.3.2   Parameter estimation

The parameters of a GARCH(1,1) can be estimated by maximizing the log likelihood

$$\log L(\alpha_0, \alpha, \lambda | r_1, ..., r_N) = -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \sum_{t=1}^{N} \left(\ln \sigma_t^2 + \frac{r_t^2}{\sigma_t^2} \right)$$

Many statistical packages now have an option for estimating GARCH models, for example in the package R,[3] one can use the function `garch(sample, order = c(p,q))`. For example `garch(sample, order = c(1, 1))` fits a GARCH(1,1) to the sample. The GARCH(1,1) can also be easily fitted in a spreadsheet package such as Excel. In this case the above likelihood function can be maximized using the Excel solver.

Example 12.1   GARCH(1,1) estimates of 1-month forward natural gas contract

The GARCH(1,1) estimates for the 1-month natural gas forward shown in Figure 12.5 are

$$\tilde{\sigma}_{t+1}^2 = 0.000028 + 0.0084 r_{t-1}^2 + 0.9653 \tilde{\sigma}_{t-1}^2.$$

Volatility persistence is equal to 0.9653 with the long term volatility estimated as

$$\tilde{\sigma}^2 = \frac{0.000028}{1 - (0.9653 - 0.0084)} = 0.001065.$$

The estimate of $\lambda + \alpha = 0.9737$ and is close to 1. Given the sampling error, it is possible that the long run variance may not be well defined. Given this and as with EWMA estimators, the model is probably best suited to forecasting near term volatility.

A possible problem with the GARCH(1,1) is that the impact of the current return on the forecast of volatility is quadratic. A day with an exceptionally large absolute return, such as might be expected following news of a major development in an alternative source of energy, or political instability in a major player in the supply of natural gas, can cause instability in parameter estimation. In general the volatility persistence parameter for energy assets is high; given this, a high volatility event will result in a (possibly) inappropriate and sustained impact on forecast volatility.

12.3.3   Leverage effect GARCH models

In empirical data of price returns there appears to be an asymmetric effect in terms of the response of volatility to good and bad news. Bad news, resulting in a negative return, increases volatility by more than good news, resulting in a positive return, of the same magnitude. This is known as the leverage effect and occurs when a fall in returns is accompanied by an increase in volatility greater than the volatility induced by an increase in returns. The idea is illustrated Figure 12.10. This asymmetry is incorporated
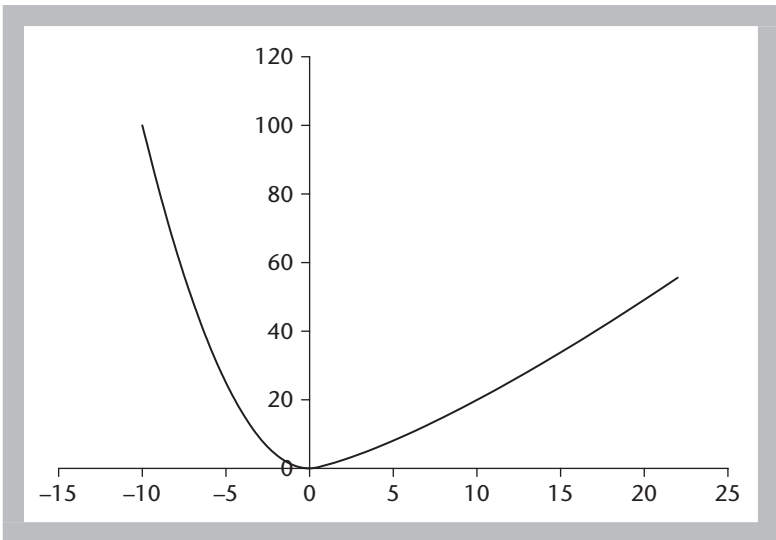
Figure 12.10 The leverage effect – how volatility reacts to bad and good news

into GARCH type models by the inclusion of a leverage effect parameter. Four popular models for capturing leverage are the Exponential GARCH model (EGARCH), Quadratic GARCH (QGARCH), Glosten-Jagannathan-Runkle and GARCH(GJR–GARCH).

The EGARCH(1,1) model takes the form

$$\ln(\tilde{\sigma}_{t+1}^2) = \alpha_0 + \alpha_1 \ln(\tilde{\sigma}_t^2) + \alpha_2 \frac{r_t}{\tilde{\sigma}_t} + \alpha_3 \frac{|r_t|}{\tilde{\sigma}_t}$$

Therefore EGARCH models the log of the conditional variance as an autoregressive moving average structure with asymmetric innovations. Since the equation uses the "log" of the process volatility it is guaranteed to be positive. The forecast of volatility is a exponential function of current volatility and product return. The parameter α_2 captures the leverage effect . A non-zero values indicates an asymmetric effect in volatility, with a negative value indicating that bad news has a larger impact on volatility than good news. A drawback of this specification is that the future volatility beyond one period ahead cannot be forecast analytically.

The QGARCH(1,1) model is given by

$$\ln(\tilde{\sigma}_{t+1}^2) = \alpha_0 + \alpha_1 r_t^2 + \alpha_2 \tilde{\sigma}_t^2 + \alpha_3 r_t.$$

The parameter α_3 captures the degree of leverage. An alternative specification uses an indicator or threshold parameter to capture leverage. Typical of this class of models is the GJR–GARCH volatility model. The GJR–GARCH(p, q) model takes the form

$$\tilde{\sigma}_{t+1}^2 = \alpha_0 + \sum_{i=1}^{q}(\alpha_i + \theta I_{t-i+1})r_t^2 + \sum_{i=1}^{p} \beta_i \tilde{\sigma}_{t-i+1}^2,$$

where I_t is the indicator variable which equals 1 if the current period return is negative and zero if the current period return is positive. Leverage is captured in the parameter $\alpha_3 > 0$.

12.3.4   Explanatory variable GARCH models

As with the regression models discussed in Chapter 8, GARCH models can include explanatory variables. For example, in modeling the volatility on a natural gas forward we might wish to include the implied volatility from an option contract on the same commodity or the volume of trade. Given k

explanatory variable $\{X_1, ..., X_k\}$ we could write the GARCH(p, q) model as

$$\tilde{\sigma}_{t+1}^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i r_{t+1-i}^2 + \sum_{j=1}^{p} \lambda_j \tilde{\sigma}_{t+1-j}^2 + \sum_{i=1}^{k} \theta_i X_i.$$

As an example consider the GARCH(1,1) with the implied option price volatility of the underlying derivatives contract (denoted by VIX) used as an explanatory variable

$$\tilde{\sigma}_{t+1}^2 = \alpha_0 + \alpha r_t^2 + \lambda \tilde{\sigma}_t^2 + \theta_1 \text{VIX}$$

Given the possibility of a leverage effect we might also consider an EGARCH(1,1)

$$\ln(\tilde{\sigma}_{t+1}^2) = \alpha_0 + \alpha_1 \ln(\tilde{\sigma}_t^2) + \alpha_2 \frac{r_t}{\tilde{\sigma}_t} + \alpha_3 \frac{|r_t|}{\tilde{\sigma}_t} + \theta_1 \text{VIX},$$

or an QGARCH(1,1)

$$\ln(\tilde{\sigma}_{t+1}^2) = \alpha_0 + \alpha_1 r_t^2 + \alpha_2 \tilde{\sigma}_t^2 + \alpha_3 r_t + \theta_1 \text{VIX}.$$

Of course there are many other GARCH models. Details of some of these are given in the section on further reading at the end of this chapter.

12.4 SUMMARY

The volatility of energy product price returns is time-varying and predictable. However, forecasting the future level of volatility is difficult because volatility forecasts are sensitive to the type and particular specification of the volatility model used. In this chapter we have given two practical approaches to building such models, one based on EWMA and the other on GARCH. In choosing which approach to adopt it is important to strike the right balance between capturing the salient feature of the data and model parsimony.

12.5 FURTHER READING

Further discussion of EWMA and REWMA type models, with particular emphasis on its role in value at risk calculations can be found in:

Guermat, C. and Harris, R. D. F. (2001) "Robust conditional variance estimation and value-at-risk," Discussion Papers in Accounting and Finance, 01/06, University of Exeter.

More general reading on the theoretical development and various applications of GARCH models is given in:

Akgiray, V. (1989) "Conditional heteroskedasticity in time series of stock returns: evidence and forecasts," *Journal of Business*, 62, 55–80.

Andersen, T. G. (1994) "Stochastic autoregressive volatility: a framework for volatility modeling," *Mathematical Finance*, 4, 75–102.

Andersen, T. G. and Bollerslev, T. (1997) "Heterogeneous information arrivals and return volatility dynamics: uncovering the long-run in high frequency returns," *Journal of Finance*, 52, 975–1005.

Baillie, R. T., Bollerslev, T., and Mikkelsen, H. O. (1996) "Fractionally integrated generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, 74, 3–30.

Ball, C. A. and Torous, W. N. (1984) "The maximum likelihood estimation of security price volatility: theory, evidence and applications to option pricing," *Journal of Business*, 57, 97–112.

Bollerslev, T. (1986) "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, 31, 307–27.

Bollerslev, T. (1987) "A conditional heteroskedastic time series model for speculative prices and rates of return," *Review of Economics and Statistics*, 69, 542–47.

Bollerslev, T., Engle, R. F., and Nelson, D. B. (1994) "ARCH Models," in *Handbook of Econometrics* Vol. IV (R. F. Engle and D. McFadden, eds) North Holland Press, Amsterdam.

Bollerslev, T. and Mikkelsen, H. O. (1996) "Modeling and pricing long-memory in stock market volatility," *Journal of Econometrics*, 73, 151–84.

Bollerslev, T. and Wooldridge, J. M. (1992) "Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances," *Econometric Reviews*, 11, 143–72.

Boudoukh, J., Richardson, M., and Whitelaw, R. F. (1997) "Investigation of a class of volatility estimators," *Journal of Derivatives*, Spring, 2(3) 63–71.

Brailsford, T. J. and Faff, R. W. (1996) "An evaluation of volatility forecasting techniques," *Journal of Banking and Finance*, 20, 419–38.

Cumby, R., Figlewski, S., and Hasbrouck, J. (1993) "Forecasting volatility and correlations with EGARCH models," *Journal of Derivatives*, Winter, 51–63.

Day, T. E. and Lewis, C. M. (1992) "Stock market volatility and the information content of stock index options," *Journal of Econometrics*, 52, 267–87.

Diebold, F. X. and Mariano, R. S. (1995) "Comparing predictive accuracy," *Journal of Business and Economic Statistics*, 13, 253–63.

Dimson, E. and Marsh, P. (1990) "Volatility forecast without data-snooping," *Journal of Banking and Finance*, 14, 399–421.

Drost, F. C. and Nijman, T. E. (1993) "Temporal aggregation of GARCH processes," *Econometrica*, 61, 909–27.

Drost, F. C. and Werker, B. J. M. (1996) "Closing the GARCH gap: continuous time GARCH modeling," *Journal of Econometrics*, 74, 31–57.

Engle, R. F. (1982) "Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation," *Econometrica*, 50, 987–1008.

Engle, R. F. and Bollerslev, T. (1986) "Modelling the persistence of conditional variances," *Econometric Reviews*, 5, 1–50.

Figlewski, S. (1997) "Forecasting volatility," *Financial Markets, Institutions and Instruments*, 6, 1–88.

Frennberg, P. and Hansson, B. (1996) "An evaluation of alternative models for predicting stock volatility: evidence from a small stock market," *Journal of International Financial Markets, Institutions and Money*, 5, 117–34.

Garman, M. B. and Klass, M. J. (1980) "On the estimation of security price volatilities from historical data," *Journal of Business*, 53, 67–78.

Goodhart, Charles A. E. and Maureen O'Hara (1997) "High frequency data in financial markets: issues and applications," *Journal of Empirical Finance*, 4, 73–114.

Heynen, R. C. and Kat, H. M. (1994) "Volatility prediction: a comparison of the stochastic volatility, GARCH(1,1), and EGARCH(1,1) models," *Journal of Derivatives*, Winter, 50–65.

Hsieh, D. A. (1993) "Implications of nonlinear dynamics for financial risk management," *Journal of Financial and Quantitative Analysis*, 28, 41–64.

Lamoureux, C. G. and Lastrapes, W. D. (1993) "Forecasting stock returns variances: towards understanding stochastic implied volatility," *Review of Financial Studies*, 6, 293–326.

Lopez, J. A. (2001) "Evaluating the predictive accuracy of volatility models," *Journal of Forecasting*, 20, 87–109.

Mandelbrot, B. B. (1963) "The variation of certain speculative prices," *Journal of Business*, 36, 394–419.

Pagan, A. R. and Sabau, H. (1992) "Consistency tests for heteroskedasticity and risk models," *Estudios Económicos*, 7, 3–30.

Rogers, L. C. G. and Satchell, S. E. (1991) "Estimating variance from high, low and closing prices," *Annals of Applied Probability*, 1, 504–12.

Schwert, G. W. (1989) "Why does stock market volatility change over time?," *Journal of Finance*, 44, 1115–53.

12.6   REVIEW QUESTIONS

1 Discuss the relative merits of EWMA estimators.

2 Contrast the EWMA with the GARCH(1,1) model – what do they have in common?

3 Why do you feel GARCH models are popular?

4 Explain how to estimate the parameters of a GARCH(1,1). How would you obtain the estimates for a GARCH(p, q)?

5 Why might it be preferable to use a GARCH(p, q) instead of a GARCH(1,1)?

6 Explain the notion of a leverage effect. Is this relevant for energy products?

Stochastic Differential Equations for Derivative Pricing and Energy Risk Management

The main objective of this chapter is twofold. First we introduce a number common stochastic processes used in the valuation of derivative contracts and financial simulations. Second, we consider their relevance to energy risk modeling.

13.1 WHAT ARE STOCHASTIC DIFFERENTIAL EQUATIONS?

Stochastic differential equations are simply probability models that allow us to model price evolution through time, and assign probabilities to possible prices as a function of the current spot and forward prices. In valuing derivative contracts, capturing the salient features of the underlying spot asset is often more important than the forecasting ability of the stochastic model. In other words a precursor for reliable valuation of derivative contracts is that the proposed stochastic process offers an accurate description of the underlying price process.

13.1.1 The key process – Geometric Brownian Motion (GBM)

A common basic assumption in the wider finance community is that financial prices change smoothly based on the premise that markets tend to anticipate

the arrival of information and process it continuously. The key stochastic differential equation used to capture this idea is the GBM

$$dS_t = \mu S_t dt + \sigma S_t dz_t, \quad \text{where } dz = \varepsilon_t \sqrt{dt} \quad \text{and} \quad \varepsilon_t \sim N(0,1).$$

S_t is the price at time t and ε_t is the source of the uncertainty. The above equation is short hand notation for the actual stochastic differential equation to describe the evolution of S_t given by

$$S_t = S_0 + \int_{u=0}^{t} \mu S_u du + \int_{u=0}^{t} \sigma S_u dz_u.$$

The first integral on the right hand side is a regular Riemann integral, the second integral is a stochastic integral requiring the tools of stochastic calculus to evaluate. We need not concern ourselves with the details of stochastic calculus at this stage, we simply need to note that it is the stochastic integral which drives the uncertainty in the model.

We can rewrite the short hand version of the model as

$$\frac{dS_t}{S_t} = \mu \, dt + \sigma \, dz_t.$$

The parameter μ is known as the instantaneous mean of the process. We can think of it as the percentage drift in the process. For example, if $\mu = 0.1$ then, if the stochastic element σdz_t was not present, the energy product would rise by 10% in a year. The second term σdz_t is the stochastic or uncertain component where the parameter σ is the process volatility. If $\sigma = 0.3$ then approximately one standard deviation over a year would be 30% of the present price of the asset.

Geometric Brownian Motion implies that prices have a lognormal distribution which in turn implies that the continuously compounded returns $\ln[S_{t+k}/S_t](k > 0)$ follows the normal distribution with mean μt and variance $\sigma^2 t$. This idea is illustrated in Figure 13.1, where the upper diagram shows a sample path of GBM, the middle diagram is a histogram of the S_t and the bottom diagram shows the distribution of $\ln(S_t)$ which, as expected, is normally distributed.

13.1.2   GBM for energy prices – is it relevant?

The popularity of GBM lies in its compatibility with the efficient market hypothesis of financial economics and in the fact that it often leads to analytically tractable solutions to derivative pricing problems. For example, the assumption of GBM is essential to derive the Black–Scholes option pricing
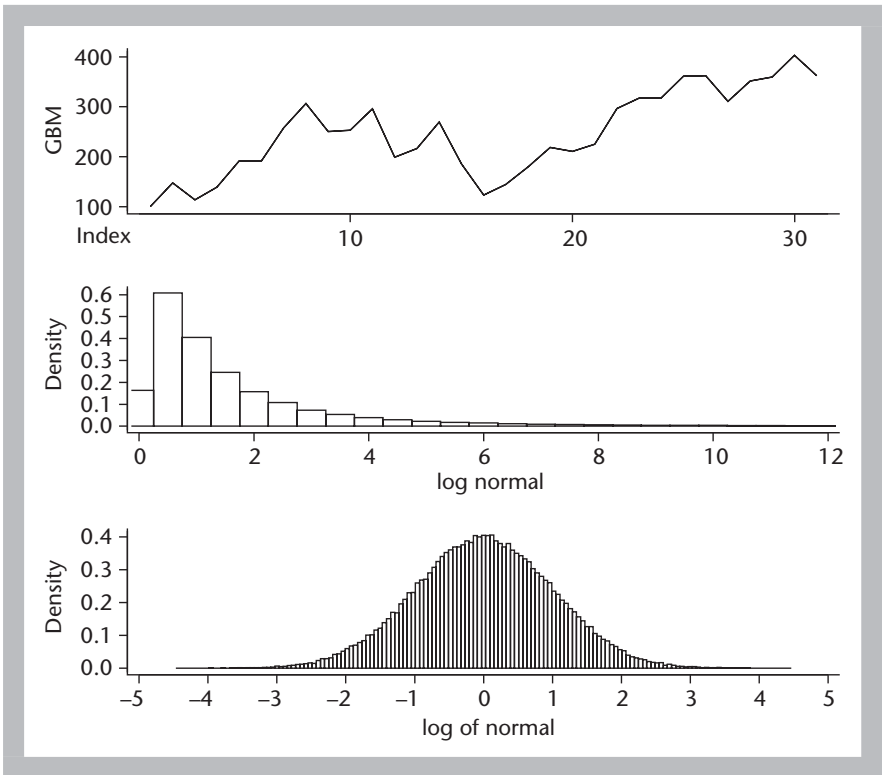
Figure 13.1 Sample path of simulated GBM price (top), histogram of the price distribution (middle), and the distribution of the log values (bottom)

formula. The question is whether such a process is appropriate for these and other energy products. One way to asses this is to investigate whether the distribution of energy prices is log normal or alternatively whether the log of prices is normally distributed. Figure 13.2 shows the Nord pool Electricity average spot price, Brent Crude price, and Natural Gas price over the period December 1999 to December 2004. The price evolution is characterized by periods of relative claim interspersed with sudden upward bursts in price and just as rapid declines.

Figures 13.3, 13.4, and 13.5 give the probability plots for the Natural Gas, Brent Crude, and Nord pool Electricity price respectively.[1] In all cases the data depart from the fitted line most evidently in the extremes, or distribution tails. This departure is particularly marked in the case of electricity where there is a clear tendency for the returns to be heaver in both tails than a lognormal distribution because the smallest points are above the line and the largest points are just below the line. We test this assumption of lognormality formally and find that the Anderson–Darling test statistic of all
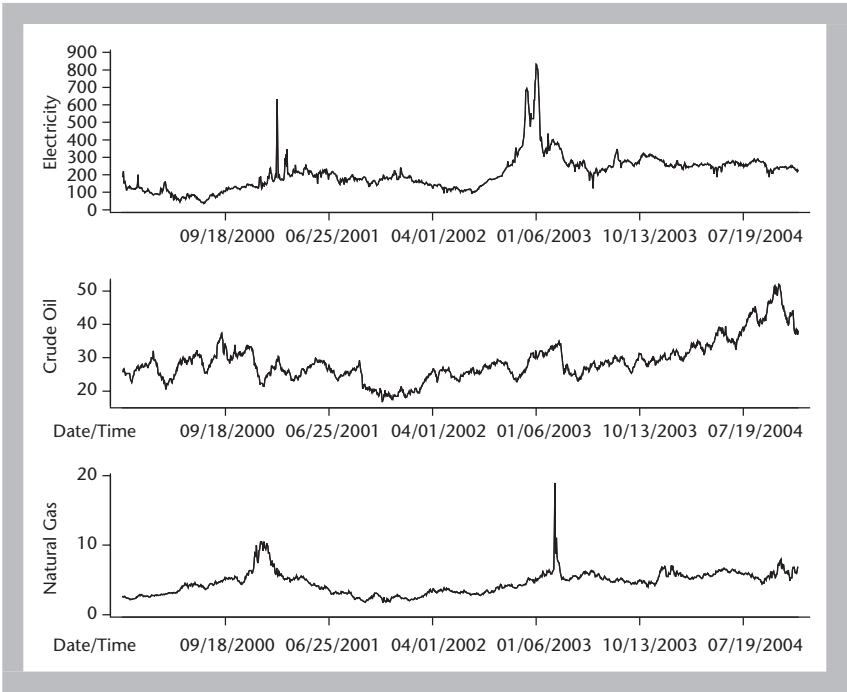
Figure 13.2 Spot price of Nord pool Electricity average spot price, Brent Crude price, and Natural Gas price over the period December 1999 to December 2004
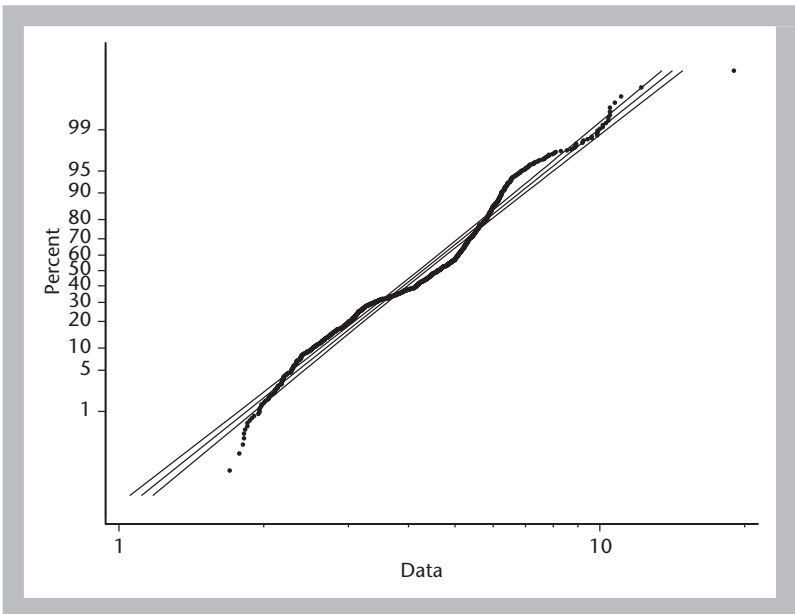
Figure 13.3 Probability plot for the Natural Gas spot price
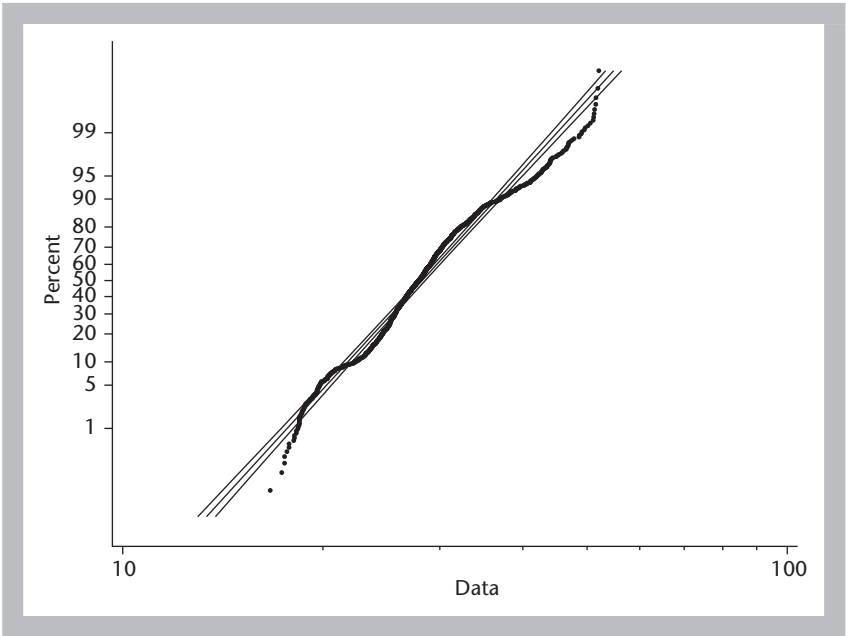
Figure 13.4 Probability plot for the Brent Crude spot price
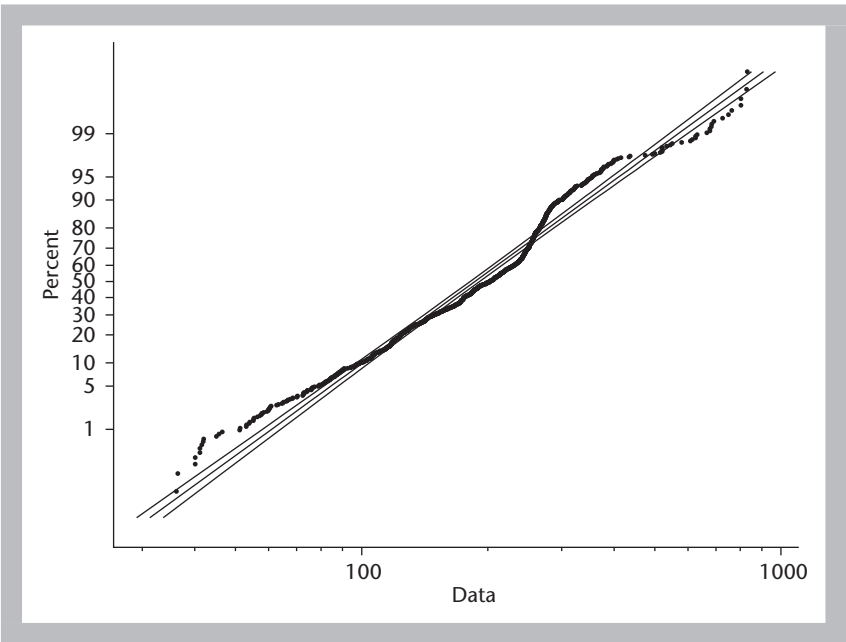
Figure 13.5 Probability plot for the Nord pool average electricity spot price

three commodities indicates significant departure from the null hypothesis of lognormalilty.

The usefulness of the GBM as a theoretical tool in general finance is well established. Its adequacy for energy price modeling may be questionable on empirical grounds. It may also be questionable on theoretical grounds because assuming that a variable changes continuously, as for the GBM, precludes large changes, or jumps, which may result from the sudden arrival of "lumpy" information.[2] This assumption may be unreasonably restrictive for many energy products, especially non-storable products such as electricity.

13.2 DEALING WITH JUMPS IN ENERGY PRICES

Energy prices often exhibit sudden, unexpected, and discontinuous changes. The presence of jumps is a significant component of the behavior of the United Kingdom electricity (PX) spot price illustrated in Figure 13.6. In fact the empirical characteristic of sudden jumps in many energy products has been invoked to explain the discrepancies observed between actual pricing of energy exotic options and theoretical predictions. More generally, ignoring heavy tails results in extreme price changes being underestimated and could lead to ineffective hedging strategies and/or to mis-pricing assets.[3] A sensible stochastic process for energy products should capture
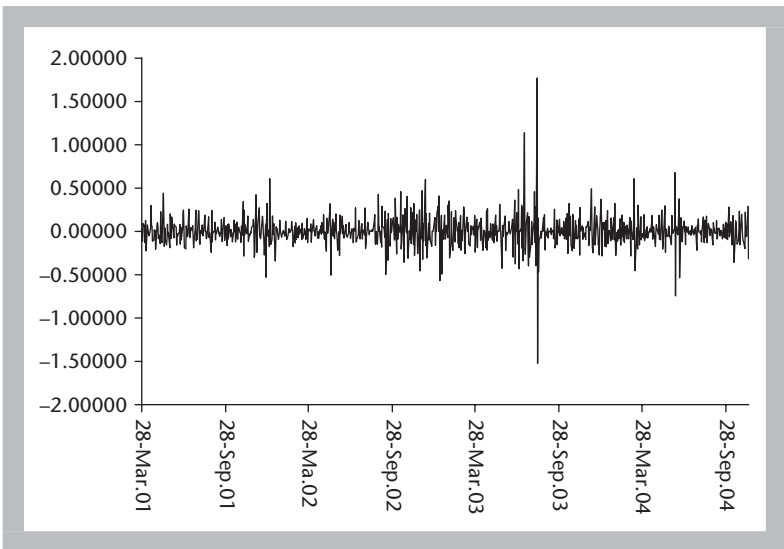
Figure 13.6 Daily log return of the UKPX-Electricity (Spot £/M) over the period March 28, 2001 to November 16, 2004

the consequence of temporary supply shocks[4] which result in price jumps. Merton (1976) introduced a model which captures this idea. It describes the evolution of the asset price by a continuous diffusion part and a discontinuous jump part.

13.2.1   Merton's Jump Model (MJM)

The MJM takes the form

$$\frac{dS_t}{S_t} = \mu\,dt + \sigma\,dz_t + J_t dq_t$$

where dq_t is a Poisson random[5] variable that takes the value zero (no jump) with probability $1 - \lambda dt$ and the value one (if jump occurs) with probability λdt. The parameter q counts the number of jumps that have occurred and J_t represents the magnitude of the jump it is assumed to be independently identically log-normally distributed with parameters (α, β^2). Since the jumps are driven by an independently identically distributed Poisson process they are entirely independent of one another. The jump process dq is a discrete time process in the sense that jumps do not occur continuously but at specific instants of time.

How does the use of this type of model compare with GBM? We see that in MJM, the returns process consists of three components, a linear drift μdt, a Brownian motion σdz representing "normal" price fluctuations that are due to normal clearing imbalances in demand and supply, and a compound Poisson process dq that accounts for jumps in prices due to the arrival of "news." Thus the continuous part of the stochastic differential equation $\mu dt + \sigma dz$ accounts for the usual fluctuations in S_t, and the jump part dq, driven by a Poisson process, accounts for the extreme events.[6] In actual fact in any small time interval, say Δ_t the probability that a single jump occurs is $\lambda\Delta_t$ whilst the overall probability that exactly q jumps occur is $(\lambda t)^q \exp(-\lambda t)/q!$ and the average waiting time for a jump is equal to $1/\lambda$. Typically, for most of the time $dq = 0$ and the process behaves as GBM.

On one level and intuitively, we may feel more comfortable with an MJM type model because by introducing jumps, we can better capture the abrupt market price changes driven by unpredictable events such as abnormal weather or forced capacity outages. More formally it can be shown that the k period returns under GBM have both skew and relative kurtosis of zero. However, under the MJM the skew is given by

$$\frac{\lambda\alpha^3}{(\sigma^2 + \lambda(\beta^2 + \alpha^2))^{1.5}\sqrt{k}},$$

and the relative kurtosis by

$$\frac{\lambda(\alpha^4 + 3\beta^4)}{(\sigma^2 + \lambda(\beta^2 + \alpha^2))^2 k}.$$

The MJM model has gained considerable popularity in the wider finance community because it can lead to analytical solutions for call and put options and interest rate derivatives such as caps, floors, and swaptions. However, it does not have very desirable properties in terms of analytical tractability in the case of path dependent (exotic) options or econometric estimation. In addition since there is only one jump component "good news" and "bad news" are not distinguished by their intensity or distributional characteristics. We saw in Chapter 12 that in empirical data of energy price returns there appears to be an asymmetric effect in terms of the response to good and bad news. Bad news, resulting in a negative return, increasing volatility by more than good news, resulting in a positive return, of the same magnitude. One way to incorporate this would be to have J_t drawn from a normal distribution but with different jump volatilities for "good" and "bad" news. Such an extension, although relatively easy to implement leads to a loss of analytical tractability of MJM. Two alternatives models which improve on this situation and are gaining widespread popularity in the wider finance community are the Pareto–Beta jump diffusion model of Ramezani and Zeng (1998) and the double exponential jump diffusion model of Kou (2002).

13.2.2   The Ramezani and Zeng Pareto–Beta jump diffusion model

The Ramezani and Zeng model (RZM) decomposes the evolution of asset prices into four components, a deterministic linear trend plus three independent stochastic components

$$\frac{dS_t}{S_t} = \mu dt + \sigma dz_t + \sum_{i=u,\,d} (J_t^i - 1)\, dq_t^i,$$

where

$$q_t^i \sim \begin{cases} \text{Poisson}(\lambda^u), & \text{if } i = u, \\ \text{Poisson}(\lambda^d), & \text{if } i = d, \end{cases}$$

and the jump magnitude J_t^u is from the Pareto distribution and J_t^d is from the standard Beta distribution so that

$$J_t^i \sim \begin{cases} \text{Pareto}(\alpha_u), & \text{if } i = u, \\ \text{Beta}(\alpha_d, 1), & \text{if } i = d. \end{cases}$$

The first stochastic component $\sigma\,dz$ represents "normal" price fluctuations. the second source of randomness is due to the arrival of good news and is driven by a Poisson process which leads to "abnormal" upward movements in price. The third source of randomness is due the arrival of bad news and is driven by a standard Beta process which leads to "abnormal" downward movements in price.

13.2.3   Kou's double exponential jump diffusion model

The double exponential jump diffusion model of Kou differs from RZM in that (J_t^i) is assumed to be a sequence of independently identically distributed nonnegative random variables such that the natural logarithm given by $\ln(J_t^i)$ has an asymmetric double exponential distribution thus

$$\ln(J_t^i) \sim \begin{cases} \xi^u \sim \text{exponential}(\eta_1) = \eta_1 \exp\{-\eta_1 \ln(J_t^u)\} \text{with probability } p \\ \xi^d \sim \text{exponential}(\eta_2) = \eta_2 \exp\{-\eta_2 \ln(J_t^d)\} \text{with probability } q \end{cases}$$

where $q = 1 - p$

In this case ξ^u and ξ^d are exponential random variables with means $1/\eta_1$ and $1/\eta_2$, respectively.

Ramezani and Zeng/Kou type models are beginning to gain popularity partly because their distribution of price returns is asymmetric and fat tailed and partly because these models lead to nearly analytical solutions to many option pricing problems. They therefore offer some advantages over MJM which is largely confined to pricing plain vanilla European options.

13.3   MODELING MEAN REVERSION

We have already seen that GBM assumes that the variance of the distribution of the price return grows linearly with time. In other words, the further out in time we gaze, the greater is our uncertainty about the value the factor will take. However, for commodities, such as crude oil and coal, we would expect supply shocks to be reasonably short lived with prices in general fluctuating around values determined by the cost of production and changes in demand.

We might therefore reasonably expect energy prices to revert back to their long term mean. Such behavior is known as mean reversion.

A simple way to model mean reversion is via an Ornstein–Uhlenbeck Process (OUP)

$$dS_t = \alpha(S^* - \ln(S_t)) \, dt + \sigma \, dz_t.$$

In this model, the asset price mean reverts to the long term level S^* at a speed given by the strictly positive parameter α – known as the mean reversion rate. An illustration will help illuminate this point. Suppose $\ln(S_t)$ represents the spread between the price at Henry Hub and the price at Houston Ship Channel. If the price at Henry Hub is greater than the price at Houston Ship Channel plus transportation costs between the two locations then the spread $\ln(S_t)$ will be above the long term level S^* and the linear drift term $\alpha(S^* - \ln(S_t))dt$ will be negative. In this situation market players will buy at Houston Ship Channel and sell at Henry Hub. This will reduce demand at Henry Hub and simultaneously drive up demand at Houston Ship Channel. Thereby placing downward pressure on the price at Henry Hub and the spread will move back towards the long term level. Similarly, if the spread is below the long term level S^* the linear drift term $\alpha(S^* - \ln(S_t))dt$ will be positive and the spread will gradually move back towards the long term level. The rate of adjustment back toward the long term level is determined by the mean reversion parameter α. The larger α, the quicker the price mean reverts.

Parameters can be estimated using the linear regression model discussed in Chapter 8. To see that this is so we rewrite the model in a discretized form

$$\Delta S_t = \alpha(S^* - \ln(S_t)) \, \Delta t + \sigma \, \Delta z_t = \alpha S^* \Delta t - \alpha \Delta t \ln(S_t) + \sigma \, \Delta z_t,$$

which after a little manipulation can be written as

$$\Delta S_t = A + B \ln(S_t) + \breve{\varepsilon}_t,$$

where

$$B = -\alpha \Delta t, \quad A = \alpha S^* \Delta t, \quad \text{and} \quad \breve{\varepsilon} = \sigma \Delta z_t.$$

The parameters A and B can then be estimated as the coefficients in a simple linear regression of $\ln(S_t - S_{t-1})$ on $\ln S_t$.

The mean reversion model has experienced considerable empirical success particularly in natural gas and crude oil products. Table 13.1 give the mean reversion and regression parameter estimates for the spot price of Henry Hub, Brent Crude, and Nord pool Electricity over the period

Table 13.1 Mean reversion and regression parameter estimates for the spot price of Henry Hub, Brent Crude, and Nord pool Electricity over the period December 1999 to December 2004

Parameter	Natural Gas (Henry Hub)	Brent Crude	Electricity (Nord pool)
A	−0.0128	−0.0261	−0.07206
B	0.0094	0.0079	0.013814
α	8.88	7.50	13.1094
S*	3.94	27.25	184.3106

December 1999–December 2004. The large value of mean reversion associated with the Nordpool electricity spot is particularly striking. This value reflects the excessively "spiky" behavior of electricity prices.

13.3.1   Mean reversion with jumps

Electricity exhibits price spikes, rather than pure price jumps . That is the price does not jump to a new level and stay there instead the price reaches a high point and then declines rapidly back to its previous trading range. This characteristic can be captured by combining mean reversion and jumps into the same model. For example using MJM as the jump model and OUP for mean reversion we could specify

$$dS_t = \alpha(S^* - \ln(S_t))\,dt + \sigma\,dz_t + J_t dq_t.$$

13.4   INTRODUCING STOCHASTIC VOLATILITY INTO ENERGY PRICES

We have already seen in Chapter 13 that the assumption of a constant volatility parameter is unrealistically restrictive. Indeed, the GBM assumption of a lognormal distribution fails to explain the existence of fat tails, the asymmetry observed in the energy price-return distribution or the observation of relatively volatile periods followed by periods of low volatility. Many different theories have been suggested to deal with the idea of nonconstant volatility; for example, we discussed the application of GARCH type models in Chapter 13, and the use of mixing distributions in Chapter 5. Another alternative is to directly specify a stochastic differential equation with a time varying volatility parameter. In this approach the asset price S

satisfies the following stochastic differential equation

$$\frac{dS_t}{S_t} = \alpha \, dt + \sigma_t \, dz_t,$$

where $\sigma_t = f(Y_t)$ and $dY_t = (a + b\,Y_t)dt + cY_t\,dv_t$, and $dv_t = \rho\,dz_t + \sqrt{1 - \rho^2}\,dw_t$, where $dw_t = \tilde{\varepsilon}_t\sqrt{dt}$ and $\tilde{\varepsilon}_t \sim N(0,1)$; $dz_t = \varepsilon_t\sqrt{dt}$ and $\varepsilon_t \sim N(0,1)$.

In this set-up the model constants are given by the parameters α, a, b, and c. The random variables ε_t and $\tilde{\varepsilon}_t$ are independent standard normal random variables. This implies that Z and W are uncorrelated. The parameter ρ is the correlation between Z and V. In other words random shocks to the process variance are correlated to the random shocks to the asset price. In this model there are two sources of risk, namely the future path of the asset price and the future path of volatility.

13.4.1   Models for stochastic volatility

A wide variety of models can be expressed in this general frame work. One of the most popular is that proposed by Hull and White (1987) in which $\sigma_t = \sqrt{Y_t}$. In The Hull and White model high volatility of the volatility parameter drives the fat tails of the price distribution. Thus extreme returns of positive and negative sign are more likely than in GBM where the asset price follows a lognormal distribution. If $\rho = 0$ so that shocks to returns and shocks to volatility are uncorrelated the price distribution is symmetric and leptokurtic. In fact the sign of the correlation determines the symmetry of the distribution; negative correlation results in the left tail of the price distribution containing more probability mass than the right tail.

Scott (1987) considered the case where the logarithm of the volatility is an OUP and $\sigma_t = \sqrt{Y_t}$ so that $dY_t = (a - bY_t)Y_t dt + cY_t\,dv_t$. Scott also introduced another variation, which has gained some popularity where $dY_t = b(a - Y_t)dt + c\,dv_t$. Although this model can result in the volatility process becoming negative this need not be a major problem since we could in principle define volatility as the positive square root of the process Y_t.

13.4.2   Stochastic volatility or GARCH?

The question that naturally arises is whether to use stochastic volatility models or GARCH to capture the time variation in volatility of energy products. GARCH models have thus far been the most frequently applied class of time-varying volatility models in empirical research. This is mainly due to the problems which arise as a consequence of the intractability of the likelihood function of the stochastic volatility model which prohibits

its direct evaluation. However, in recent years considerable progress has been made in this area and techniques for estimation model parameters are becoming more widely available. In practice it is highly likely that the performance of the GARCH and stochastic volatility models depends on the specific product in question.[7]

13.5   SUMMARY

Almost every aspect of modern financial risk management, from value-at-risk calculations to option pricing, critically depend upon the form of the probability distribution describing the dynamics of energy prices or returns. Stochastic differential equations play a central role in describing the dynamics of the price and return processes because they allow us to model the evolution of prices through time and assign probabilities to possible future prices. These future prices can be used in derivative pricing and risk management models. In order to accurately represent the empirical characteristics of energy products the basic GBM model needs to be augmented to cater for stochastically driven jumps, mean reversion, and time varying volatility.

13.6   FURTHER READING

Merton (1976, 1982) first suggested that the returns process of financial assets consisted of the three components of linear drift, Geometric Brownian Motion representing normal price variations, and a compound Poisson process that accounts for "abnormal" changes (i.e. jumps). Further details can be found in his references listed below.

An interesting study of the effectiveness of different stochastic models used for describing the evolution of electricity spot prices (California, Scandinavia, England, and Wales) was carried out by Johnson and Barz (1999). They evaluated the effectiveness of four different stochastic models (Brownian motion, Mean Reversion, Geometric Brownian Motion, and Geometric Mean Reversion). The models were tested with and without jumps. The objective was to reproduce electricity price behavior for three distinct periods (April–May, August–September, and November–December). The authors concluded that the Geometric Mean Reverting model gave the overall best performance and the addition of jumps to any of the of the models also improved that model's performance.

The literature on Ramezani *et al*. and Kou type jump diffusion processes is large and expanding. Ramezani (2004) conducts estimation and empirical assessment of this class of models against a standard Merton type jump diffusion model and various GARCH type specifications. Alternative variations on the jump specification, including different distributional assumptions for the jump magnitude, time varying jump intensity, correlated jump magnitudes, deterministic volatility structure combined with log-normally distributed jumps, and models in which the asset price and volatility jump can be found in a number of publications, the most representative are Naik (1993), Anderson and Anderson (2002), and Eraker *et al*. (2003).

Andersen, L. and Andersen, J. (2000) "Jump-diffusion processes: volatility smile fitting and numerical methods for pricing," *Review of Derivatives Research*, 4, 231–62.

Barz, G. and Johnson, B. (1999) *Energy Modelling and the Management of Uncertainty*, Risk books, London.

Eraker, B., Johannes, M., and Polson, N. (2003) "The impact of jumps in returns and volatility," *Journal of Finance*, 58, 1269–1300.

Hull, J. and White, A. (1987) "The pricing of options on assets with stochastic volatilities." *Journal of Finance*, 42(2), 281–300.

Hull, J. and White, A. (1988) "An analysis of the bias in option pricing caused by a stochastic volatility." *Advances in Futures and Options Research*, 3, 29–61.

Johnson, B. and Graydon, B. (1999) *Energy Modelling and the Management of Uncertainty*, Risk books, London.

Kat, H. and Heynen, R. (1994) "Volatility prediction: a comparison of the stochastic volatility, GARCH (1,1) and EGARCH (1,1) models," *Journal of Derivatives*, 2(2), 50–65.

Kou, S. (2002) "A jump diffusion model for option pricing," *Management Science*, 48(8), 1086–1101.

Kou, S. and Wang, H. (2004) "Option pricing under a double exponential jump diffusion model," *Management Science*, 50(9), 1178–92.

Merton, R. C. (1976) "Option pricing when underlying stock returns are discontinuous," *Journal of Financial Economics*, 3, 125–44.

Merton, R. C. (1982) *Continuous-Time Finance*, Paperback edn, Blackwell Publishers Ltd., Cambridge, MA, and Oxford, UK.

Naik, V. (1993) "Option valuation and hedging strategies with jumps in the volatility of asset returns." *The Journal of Finance*, 48, 1969–84.

Ramezani, C. and Zeng, Y. (1998) "Maximum likelihood estimation of asymmetric jump diffusion processes: application to security prices." Working Paper. Department of Mathematics and Statistics, University of Missouri.

Ramezani, C. (2004) "An empirical assessment of the double exponential jump diffusion process." Finance Department. Orfalea College of Business. California Polytechnic.

Scott, L. O. (1987) "Option pricing when the variance changes randomly: theory, estimation and an application." *Journal of Financial and Quantitative Analysis*, 22, 419–38.

13.7   REVIEW QUESTIONS

1 Under what circumstances would you recommend the use of GBM for pricing an option on an energy product?

2 How would you assess the performance of MJM relative to RZM or Kou's model?

3 In which type of energy products would you expect to successfully use OUP?

4 What are the relative advantages of stochastic volatility models over GARCH?

5 Contrast the Hull and White (1987) model against the two Scott (1987) models. Which do you feel would be more suitable for the Nord pool spot price and why?

Appendix: Statistical Tables

This appendix contains statistical tables of the common sampling distributions used in statistical inference. See Chapter 4 for examples of their use.

CUMULATIVE DISTRIBUTION FUNCTION OF THE STANDARD NORMAL DISTRIBUTION

Table A.1 shows the probability, $F(z)$ that a standard normal random variable is less than the value z. For example, the probability is 0.9986 that a standard normal random variable is less than 3.

CHI-SQUARE DISTRIBUTION

For a given probabilities α, Table A.2 shows the values of the chi-square distribution. For example, the probability is 0.05 that a chi-square random variable with 10 degrees of freedom is greater than 18.31.

STUDENT T-DISTRIBUTION

For a given probability α, Table A.3 shows the values of the student t-distribution. For example, the probability is 0.05 that a student t random variable with 10 degrees of freedom is greater than 1.812.

Table A.1 Cumulative distribution function of the standard normal distribution

z	F(z)	z	F(z)	Z	F(z)	z	F(z)
0	0.5	0.32	0.625516	0.64	0.738914	0.96	0.831472
0.01	0.503989	0.33	0.6293	0.65	0.742154	0.97	0.833977
0.02	0.507978	0.34	0.633072	0.66	0.745373	0.98	0.836457
0.03	0.511967	0.35	0.636831	0.67	0.748571	0.99	0.838913
0.04	0.515953	0.36	0.640576	0.68	0.751748	1	0.841345
0.05	0.519939	0.37	0.644309	0.69	0.754903	1.01	0.843752
0.06	0.523922	0.38	0.648027	0.7	0.758036	1.02	0.846136
0.07	0.527903	0.39	0.651732	0.71	0.761148	1.03	0.848495
0.08	0.531881	0.4	0.655422	0.72	0.764238	1.04	0.85083
0.09	0.535856	0.41	0.659097	0.73	0.767305	1.05	0.85314
0.1	0.539828	0.42	0.662757	0.74	0.77035	1.06	0.85542
0.11	0.543795	0.43	0.666402	0.75	0.773373	1.07	0.85769
0.12	0.547758	0.44	0.670031	0.76	0.776373	1.08	0.85992
0.13	0.551717	0.45	0.673645	0.77	0.77935	1.09	0.86214
0.14	0.55567	0.46	0.677242	0.78	0.782305	1.1	0.86433
0.15	0.559618	0.47	0.680822	0.79	0.785236	1.11	0.8665
0.16	0.563559	0.48	0.684386	0.8	0.788145	1.12	0.86864
0.17	0.567495	0.49	0.687933	0.81	0.79103	1.13	0.87076
0.18	0.571424	0.5	0.691462	0.82	0.793892	1.14	0.87285
0.19	0.575345	0.51	0.694974	0.83	0.796731	1.15	0.87492
0.2	0.57926	0.52	0.698468	0.84	0.799546	1.16	0.87697
0.21	0.583166	0.53	0.701944	0.85	0.802338	1.17	0.87899
0.22	0.587064	0.54	0.705402	0.86	0.805106	1.18	0.881
0.23	0.590954	0.55	0.70884	0.87	0.80785	1.19	0.88297
0.24	0.594835	0.56	0.71226	0.88	0.81057	1.2	0.88493
0.25	0.598706	0.57	0.715661	0.89	0.813267	1.21	0.88686
0.26	0.602568	0.58	0.719043	0.9	0.81594	1.22	0.88876
0.27	0.60642	0.59	0.722405	0.91	0.818589	1.23	0.89065
0.28	0.610261	0.6	0.725747	0.92	0.821214	1.24	0.89251
0.29	0.614092	0.61	0.729069	0.93	0.823814	1.25	0.89435
0.3	0.617911	0.62	0.732371	0.94	0.826391	1.26	0.89616
0.31	0.621719	0.63	0.735653	0.95	0.828944	1.27	0.89795

Continued

Table A.1 Continued

z	F(z)	z	F(z)	Z	F(z)	z	F(z)
1.28	0.89972	1.63	0.9484493	1.98	0.976148	2.33	0.990097
1.29	0.90147	1.64	0.9494972	1.99	0.976705	2.34	0.990358
1.3	0.90319	1.65	0.9505297	2	0.97725	2.35	0.990613
1.31	0.90490	1.66	0.9515438	2.01	0.977784	2.36	0.990863
1.32	0.90658	1.67	0.952546	2.02	0.978308	2.37	0.991106
1.33	0.90824	1.68	0.9535219	2.03	0.978822	2.38	0.991344
1.34	0.90987	1.69	0.954486	2.04	0.979325	2.39	0.991576
1.35	0.911492	1.7	0.9554357	2.05	0.979818	2.4	0.991802
1.36	0.913085	1.71	0.956367	2.06	0.980301	2.41	0.992024
1.37	0.914656	1.72	0.957284	2.07	0.980774	2.42	0.99224
1.38	0.916207	1.73	0.9581857	2.08	0.981237	2.43	0.992451
1.39	0.917736	1.74	0.9590711	2.09	0.981691	2.44	0.992656
1.4	0.919243	1.75	0.9599412	2.1	0.982136	2.45	0.992857
1.41	0.92073	1.76	0.960796	2.11	0.982571	2.46	0.993053
1.42	0.922196	1.77	0.9616365	2.12	0.982997	2.47	0.993244
1.43	0.923641	1.78	0.9624628	2.13	0.983414	2.48	0.993431
1.44	0.925066	1.79	0.9632737	2.14	0.983823	2.49	0.993613
1.45	0.926471	1.8	0.96407s5	2.15	0.984222	2.5	0.99379
1.46	0.927855	1.81	0.9648529	2.16	0.984614	2.51	0.993963
1.47	0.929219	1.82	0.9656212	2.17	0.984997	2.52	0.994132
1.48	0.930563	1.83	0.9663752	2.18	0.985371	2.53	0.994297
1.49	0.931888	1.84	0.9671161	2.19	0.985738	2.54	0.994457
1.5	0.933193	1.85	0.9678437	2.2	0.986097	2.55	0.994614
1.51	0.934478	1.86	0.968557	2.21	0.986447	2.56	0.994766
1.52	0.935744	1.87	0.969258	2.22	0.986791	2.57	0.994915
1.53	0.936992	1.88	0.969946	2.23	0.987126	2.58	0.99506
1.54	0.93822	1.89	0.970621	2.24	0.987455	2.59	0.995201
1.55	0.939429	1.9	0.971284	2.25	0.987776	2.6	0.995339
1.56	0.940621	1.91	0.971933	2.26	0.988089	2.61	0.995473
1.57	0.9417928	1.92	0.972571	2.27	0.988396	2.62	0.995603
1.58	0.942947	1.93	0.973197	2.28	0.988696	2.63	0.995731
1.59	0.9440839	1.94	0.97381	2.29	0.988989	2.64	0.995855
1.6	0.9452013	1.95	0.974412	2.3	0.989276	2.65	0.995975
1.61	0.9463014	1.96	0.975002	2.31	0.989556	2.66	0.996093
1.62	0.947384	1.97	0.975581	2.32	0.98983	2.67	0.996207

Continued

Table A.1 Continued

z	F(z)	z	F(z)	Z	F(z)	z	F(z)
2.68	0.996319	3.03	0.998777	3.38	0.999638	3.73	0.999904
2.69	0.996427	3.04	0.998817	3.39	0.99965	3.74	0.999908
2.7	0.996533	3.05	0.998856	3.4	0.999663	3.75	0.999912
2.71	0.996636	3.06	0.998893	3.41	0.999675	3.76	0.999915
2.72	0.996736	3.07	0.99893	3.42	0.999687	3.77	0.999918
2.73	0.996833	3.08	0.998965	3.43	0.999698	3.78	0.999922
2.74	0.996928	3.09	0.998999	3.44	0.999709	3.79	0.999925
2.75	0.99702	3.1	0.999032	3.45	0.99972	3.8	0.999928
2.76	0.99711	3.11	0.999064	3.46	0.99973	3.81	0.99993
2.77	0.997197	3.12	0.999096	3.47	0.99974	3.82	0.999933
2.78	0.997282	3.13	0.999126	3.48	0.999749	3.83	0.999936
2.79	0.997365	3.14	0.999155	3.49	0.999758	3.84	0.999938
2.8	0.997445	3.15	0.999184	3.5	0.999767	3.85	0.999941
2.81	0.997523	3.16	0.999211	3.51	0.999776	3.86	0.999943
2.82	0.997599	3.17	0.999238	3.52	0.999784	3.87	0.999946
2.83	0.997673	3.18	0.999264	3.53	0.999792	3.88	0.999948
2.84	0.997744	3.19	0.999289	3.54	0.9998	3.89	0.99995
2.85	0.997814	3.2	0.999313	3.55	0.999807	3.9	0.999952
2.86	0.997882	3.21	0.999336	3.56	0.999815	3.91	0.999954
2.87	0.997948	3.22	0.999359	3.57	0.999821	3.92	0.999956
2.88	0.998012	3.23	0.999381	3.58	0.999828	3.93	0.999958
2.89	0.998074	3.24	0.999402	3.59	0.999835	3.94	0.999959
2.9	0.998134	3.25	0.999423	3.6	0.999841	3.95	0.999961
2.91	0.998193	3.26	0.999443	3.61	0.999847	3.96	0.999963
2.92	0.99825	3.27	0.999462	3.62	0.999853	3.97	0.999964
2.93	0.998305	3.28	0.999481	3.63	0.999858	3.98	0.999966
2.94	0.998359	3.29	0.999499	3.64	0.999864	3.99	0.999967
2.95	0.998411	3.3	0.999517	3.65	0.999869	4	0.999968
2.96	0.998462	3.31	0.999533	3.66	0.999874		
2.97	0.998511	3.32	0.99955	3.67	0.999879		
2.98	0.998559	3.33	0.999566	3.68	0.999883		
2.99	0.998605	3.34	0.999581	3.69	0.999888		
3	0.99865	3.35	0.999596	3.7	0.999892		
3.01	0.998694	3.36	0.99961	3.71	0.999896		
3.02	0.998736	3.37	0.999624	3.72	0.9999		

Table A.2 Cut off points for the chi-square distribution

Degrees of freedom	α				
	0.005	0.01	0.025	0.05	0.1
1	7.88	6.63	5.02	3.84	2.71
2	10.60	9.21	7.38	5.99	4.61
3	12.84	11.34	9.35	7.81	6.25
4	14.86	13.28	11.14	9.49	7.78
5	16.75	15.09	12.83	11.07	9.24
6	18.55	16.81	14.45	12.59	10.64
7	20.28	18.48	16.01	14.07	12.02
8	21.95	20.09	17.53	15.51	13.36
9	23.59	21.67	19.02	16.92	14.68
10	25.19	23.21	20.48	18.31	15.99
11	26.76	24.73	21.92	19.68	17.28
12	28.30	26.22	23.34	21.03	18.55
13	29.82	27.69	24.74	22.36	19.81
14	31.32	29.14	26.12	23.68	21.06
15	32.80	30.58	27.49	25.00	22.31
16	34.27	32.00	28.85	26.30	23.54
17	35.72	33.41	30.19	27.59	24.77
18	37.16	34.81	31.53	28.87	25.99
19	38.58	36.19	32.85	30.14	27.20
20	40.00	37.57	34.17	31.41	28.41
21	41.40	38.93	35.48	32.67	29.62
22	42.80	40.29	36.78	33.92	30.81
23	44.18	41.64	38.08	35.17	32.01
24	45.56	42.98	39.36	36.42	33.20
25	46.93	44.31	40.65	37.65	34.38
26	48.29	45.64	41.92	38.89	35.56
27	49.65	46.96	43.19	40.11	36.74
28	50.99	48.28	44.46	41.34	37.92
29	52.34	49.59	45.72	42.56	39.09
30	53.67	50.89	46.98	43.77	40.26
31	55.00	52.19	48.23	44.99	41.42
32	56.33	53.49	49.48	46.19	42.58
33	57.65	54.78	50.73	47.40	43.75
34	58.96	56.06	51.97	48.60	44.90

Continued

Table A.2 Continued

Degrees of freedom	α				
	0.005	0.01	0.025	0.05	0.1
35	60.27	57.34	53.20	49.80	46.06
36	61.58	58.62	54.44	51.00	47.21
37	62.88	59.89	55.67	52.19	48.36
38	64.18	61.16	56.90	53.38	49.51
39	65.48	62.43	58.12	54.57	50.66
40	66.77	63.69	59.34	55.76	51.81
41	68.05	64.95	60.56	56.94	52.95
42	69.34	66.21	61.78	58.12	54.09
43	70.62	67.46	62.99	59.30	55.23
44	71.89	68.71	64.20	60.48	56.37
45	73.17	69.96	65.41	61.66	57.51
46	74.44	71.20	66.62	62.83	58.64
47	75.70	72.44	67.82	64.00	59.77
48	76.97	73.68	69.02	65.17	60.91
49	78.23	74.92	70.22	66.34	62.04
50	79.49	76.15	71.42	67.50	63.17
51	80.75	77.39	72.62	68.67	64.30
52	82.00	78.62	73.81	69.83	65.42
53	83.25	79.84	75.00	70.99	66.55
54	84.50	81.07	76.19	72.15	67.67
55	85.75	82.29	77.38	73.31	68.80
56	86.99	83.51	78.57	74.47	69.92
57	88.24	84.73	79.75	75.62	71.04
58	89.48	85.95	80.94	76.78	72.16
59	90.72	87.17	82.12	77.93	73.28
60	91.95	88.38	83.30	79.08	74.40
61	93.19	89.59	84.48	80.23	75.51
62	94.42	90.80	85.65	81.38	76.63
63	95.65	92.01	86.83	82.53	77.75
64	96.88	93.22	88.00	83.68	78.86
65	98.10	94.42	89.18	84.82	79.97
66	99.33	95.63	90.35	85.96	81.09
67	100.55	96.83	91.52	87.11	82.20
68	101.78	98.03	92.69	88.25	83.31

Table A.2 Continued

Degrees of freedom	α				
	0.005	0.01	0.025	0.05	0.1
69	103.00	99.23	93.86	89.39	84.42
70	104.21	100.43	95.02	90.53	85.53
71	105.43	101.62	96.19	91.67	86.64
72	106.65	102.82	97.35	92.81	87.74
73	107.86	104.01	98.52	93.95	88.85
74	109.07	105.20	99.68	95.08	89.96
75	110.29	106.39	100.84	96.22	91.06
76	111.50	107.58	102.00	97.35	92.17
77	112.70	108.77	103.16	98.48	93.27
78	113.91	109.96	104.32	99.62	94.37
79	115.12	111.14	105.47	100.75	95.48
80	116.32	112.33	106.63	101.88	96.58
81	117.52	113.51	107.78	103.01	97.68
82	118.73	114.69	108.94	104.14	98.78
83	119.93	115.88	110.09	105.27	99.88
84	121.13	117.06	111.24	106.39	100.98
85	122.32	118.24	112.39	107.52	102.08
86	123.52	119.41	113.54	108.65	103.18
87	124.72	120.59	114.69	109.77	104.28
88	125.91	121.77	115.84	110.90	105.37
89	127.11	122.94	116.99	112.02	106.47
90	128.30	124.12	118.14	113.15	107.57
91	129.49	125.29	119.28	114.27	108.66
92	130.68	126.46	120.43	115.39	109.76
93	131.87	127.63	121.57	116.51	110.85
94	133.06	128.80	122.72	117.63	111.94
95	134.25	129.97	123.86	118.75	113.04
96	135.43	131.14	125.00	119.87	114.13
97	136.62	132.31	126.14	120.99	115.22
98	137.80	133.48	127.28	122.11	116.32
99	138.99	134.64	128.42	123.23	117.41
100	140.17	135.81	129.56	124.34	118.50

Table A.3 Cut off points for the student t-distribution

Degrees of freedom	α				
	0.005	0.01	0.025	0.05	0.1
1	63.656	31.821	12.706	6.314	3.078
2	9.925	6.965	4.303	2.920	1.886
3	5.841	4.541	3.182	2.353	1.638
4	4.604	3.747	2.776	2.132	1.533
5	4.032	3.365	2.571	2.015	1.476
6	3.707	3.143	2.447	1.943	1.440
7	3.499	2.998	2.365	1.895	1.415
8	3.355	2.896	2.306	1.860	1.397
9	3.250	2.821	2.262	1.833	1.383
10	3.169	2.764	2.228	1.812	1.372
11	3.106	2.718	2.201	1.796	1.363
12	3.055	2.681	2.179	1.782	1.356
13	3.012	2.650	2.160	1.771	1.350
14	2.977	2.624	2.145	1.761	1.345
15	2.947	2.602	2.131	1.753	1.341
16	2.921	2.583	2.120	1.746	1.337
17	2.898	2.567	2.110	1.740	1.333
18	2.878	2.552	2.101	1.734	1.330
19	2.861	2.539	2.093	1.729	1.328
20	2.845	2.528	2.086	1.725	1.325
21	2.831	2.518	2.080	1.721	1.323
22	2.819	2.508	2.074	1.717	1.321
23	2.807	2.500	2.069	1.714	1.319
24	2.797	2.492	2.064	1.711	1.318
25	2.787	2.485	2.060	1.708	1.316

Table A.3 Continued

Degrees of freedom	α				
	0.005	0.01	0.025	0.05	0.1
26	2.779	2.479	2.056	1.706	1.315
27	2.771	2.473	2.052	1.703	1.314
28	2.763	2.467	2.048	1.701	1.313
29	2.756	2.462	2.045	1.699	1.311
30	2.750	2.457	2.042	1.697	1.310
40	2.704	2.423	2.021	1.684	1.303
60	2.660	2.390	2.000	1.671	1.296
100	2.626	2.364	1.984	1.660	1.290
500	2.586	2.334	1.965	1.648	1.283
1000	2.581	2.330	1.962	1.646	1.282
∞	2.576	2.327	1.960	1.645	1.282

F-DISTRIBUTION

Tables A.4, A.5 and A.6 show, for a given probability α, the values of the F-distribution. For example, the probability is 0.05 that a $F_{v,k}$ distributed random variable, with $v = 8$ and $k = 10$, is greater than 6.0.

Table A.4 Cut off points for the F-distribution where $\alpha = 0.01$

Denominator degrees of freedom = k	Numerator degrees of freedom = v									
	1	4	8	10	20	50	100	500	1000	∞
1	4052.2	5624.3	5981.0	6055.9	6208.7	6302.3	6333.9	6359.5	6362.8	6365.6
4	21.2	16.0	14.8	14.5	14.0	13.7	13.6	13.5	13.5	13.5
8	11.3	7.0	6.0	5.8	5.4	5.1	5.0	4.9	4.9	4.9
10	10.0	6.0	5.1	4.8	4.4	4.1	4.0	3.9	3.9	3.9
20	8.1	4.4	3.6	3.4	2.9	2.6	2.5	2.4	2.4	2.4
50	7.2	3.7	2.9	2.7	2.3	1.9	1.8	1.7	1.7	1.7
100	6.9	3.5	2.7	2.5	2.1	1.7	1.6	1.5	1.4	1.4
500	6.7	3.4	2.5	2.4	1.9	1.6	1.4	1.2	1.2	1.2
1000	6.7	3.3	2.5	2.3	1.9	1.5	1.4	1.2	1.2	1.1
∞	6.6	3.3	2.5	2.3	1.9	1.5	1.4	1.2	1.1	1.0

Table A.5 Cut off points for the *F*-distribution where $\alpha = 0.05$

Denominator degrees of freedom	Numerator degrees of freedom									
	1	4	8	10	20	50	100	500	1000	∞
1	161.4	224.6	238.9	241.9	248.0	251.8	253.0	254.1	254.2	254.3
4	7.7	6.4	6.0	6.0	5.8	5.7	5.7	5.6	5.6	5.6
8	5.3	3.8	3.4	3.3	3.2	3.0	3.0	2.9	2.9	2.9
10	5.0	3.5	3.1	3.0	2.8	2.6	2.6	2.5	2.5	2.5
20	4.4	2.9	2.4	2.3	2.1	2.0	1.9	1.9	1.8	1.8
50	4.0	2.6	2.1	2.0	1.8	1.6	1.5	1.5	1.4	1.4
100	3.9	2.5	2.0	1.9	1.7	1.5	1.4	1.3	1.3	1.3
500	3.9	2.4	2.0	1.8	1.6	1.4	1.3	1.2	1.1	1.1
1000	3.9	2.4	1.9	1.8	1.6	1.4	1.3	1.1	1.1	1.1
∞	3.8	2.4	1.9	1.8	1.6	1.4	1.2	1.1	1.1	1.0

Table A.6 Cut off points for the *F*-distribution where $\alpha = 0.1$

Denominator degrees of freedom	Numerator degrees of freedom									
	1	4	8	10	20	50	100	500	1000	∞
1	39.9	55.8	59.4	60.2	61.7	62.7	63.0	63.3	63.3	63.3
4	4.5	4.1	4.0	3.9	3.8	3.8	3.8	3.8	3.8	3.8
8	3.5	2.8	2.6	2.5	2.4	2.3	2.3	2.3	2.3	2.3
10	3.3	2.6	2.4	2.3	2.2	2.1	2.1	2.1	2.1	2.1
20	3.0	2.2	2.0	1.9	1.8	1.7	1.7	1.6	1.6	1.6
50	2.8	2.1	1.8	1.7	1.6	1.4	1.4	1.3	1.3	1.3
100	2.8	2.0	1.7	1.7	1.5	1.4	1.3	1.2	1.2	1.2
500	2.7	2.0	1.7	1.6	1.4	1.3	1.2	1.1	1.1	1.1
1000	2.7	2.0	1.7	1.6	1.4	1.3	1.2	1.1	1.1	1.1
∞	2.7	1.9	1.7	1.6	1.4	1.3	1.2	1.1	1.1	1.0

Notes

1 THE STATISTICAL NATURE OF ENERGY RISK MODELING

1 Taken from page 5 of the book entitled *Multivariate Statistical Analysis* by Sam Kash Kachiga (1991) published by Radius Press (New York).
2 Also known as price risk or market price risk.
3 For example, during volatile periods such as wars like the 1990s gulf war, over-the-counter dealers can simply refuse to pick up the phone to make a price.
4 The quote is from a paper entitled "The Past, Present and Future of Energy Trading" by Ed Krapels and presented at the Oil and Gas law conference, February 19, 2004. It can be downloaded from the Energy Security Analysis, Inc. (ESAI) reading room at www.esai.com/rea_readingroom.htm
5 The wellhead price is the price of gas at the site of production. It is also known as the commodity price.

2 INTRODUCTION TO APPLIED PROBABILITY FOR ENERGY RISK MANAGEMENT

1 As a historical footnote, the normal distribution was discovered by the Huguenot refugee Abraham De Moivre as the limiting form of the discrete binomial distribution in 1733 and later formally derived by Gauss in 1809 in his "Theoria motus corporum." The wide spread applicability and importance of this probability distribution was summed up succinctly by Galton who called it the "law of frequency of error." In 1899 he wrote:

> I know scarcely anything so apt to impress the imagination as the wonderful form of the cosmic order expressed by the "Law of Frequency of Error." The law would have been personified by the Greeks and deified if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason.

4 INFERENTIAL STATISTICS METHODS FOR ENERGY RISK MANAGERS

1 *R* is a widely used statistical and numerical computational tool. It runs on all of the major operating systems, is fast, easy to use, and simple to extend. It is available for free from http://www.r-project.org/

5 MODELING AND FITTING PRICE DISTRIBUTIONS

1 Notice the coefficient of skew for electricity is 0.75 and therefore of some concern. However since the difference between the median and mean is close to zero we choose to assume that the returns are symmetric, for illustrative purposes.

2 This is the asymptotic distribution. It needs to be noted that the small-sample tail quantiles of this statistic are quite different from their asymptotic counterparts (see Deb and Sefton (1996, table 1), Urzúa (1996, table 1)). Therefore, the use of asymptotic critical values in small samples will distort the actual size of the test, and may lead to incorrect decisions in applied work. For accurate small-sample critical values see Deb and Sefton (1996) who compute 14 empirical 10% and 5% significance values for sample sizes ranging from 20 to 800 observations.

3 An alternative to the approach adopted in this section is to specify the probability distributions of the mixing distributions and then estimate the parameters using the method of maximum likelihood. The advantage of using the simple method described in this section is that it forces the user to experiment and investigate the impact of changing the values of the parameters on each of the mixing distributions. The approach is useful as a conceptual tool to aid understanding and insight. The method of maximum likelihood is described in Section 5.6.1.

4 We previously estimated the parameters of the Logistic distribution using estimators derived from a technique known as the method of moments. In actual fact a maximum likelihood or least squares estimators can also be used.

5 You can also use the spelling `optimize()` – it is exactly the same function.

6 NONPARAMETRIC DENSITY ESTIMATION FOR ENERGY PRICE RETURNS

This is part of a much longer poem found on the website http://www.readingstats. com/fourth/poetry10.htm. It is claimed that the words were written by a student called Cory Lation who was taking or had just finished a course in statistics. This is probably not the case and "Cory Lation" is likely to be the fictional name of a professor of statistics. If you wish to investigate further see: http://www-personal.buseco.monash.edu.au/~hyndman/

7 CORRELATION ANALYSIS

1 We can also calculate the product moment correlation coefficient using the in-built *R* function *cor(x,y)*:

> *cor(longley$GNP, longley$Unemployed)*

[1] 0.604261

2 This is an approximate confidence interval, but is good enough for most practical purposes.
3 Economists and Econometricians have written much about the relationship between unemployment and economic growth. Our results are only tentative – they are after all based solely on the use of correlation.
4 Eydeland, A. and Wolyniec, K. (2003) *Energy and Power Risk Management*, John Wiley and Sons, New York.
5 It is also possible that the correlation does not exist, and the value we estimate is simply noise. This could be the case if the price returns are nonstationary or if correlation is the inappropriate estimator of dependence.

8   A PRIMER IN APPLIED REGRESSION ANALYSIS

1 Of course, the assumption of linearity may not always be appropriate. In such circumstances a non-linear functional form can be specified, this is the basis of non-linear regression. We discuss this issue further in later sections of this chapter.
2 This section can be skipped by those readers who do not intend to use the *R* statistical package. The objective is to get readers familiar with the *R* statistical package. *R* is freely available at www.r-project.org/
3 We will often write an estimate using the notation that is, the estimate of β would be written as $\hat{\beta}$.

9   MULTIPLE REGRESSION AND PREDICTION

1 To follow this example given in this section you will need to install the contributed package lmtestand load this using the command library(lmtest).

10   MISSPECIFICATION TESTING

1 To perform this test you will need to install the contributed package lmtest and load it using the command: library(lmtest).
2 This is true at the 1% and 5% level typically used in academic research. Notice however, that we do reject the null hypothesis at the 10% level.
3 We have previously emphasized that statistical modeling is art embedded in science. Thus, there are often slightly different interpretations. For example on the QQ plot the points lie along the reference line except in the extreme tails. Thus if we were to take a "relaxed" view, we might not reject the assumption of normality. Evidence in favor or such a decision might be gleaned from the Shapiro Wilk normality test, which is not rejected at the 1% level.

12   MODELING ENERGY PRICE VOLATILITY

1 In this case λ was chosen to maximize the correlation between the EWMA forecast and the actual squared returns. In practice values of λ between 0.95 and 0.99 are typical for energy products.

2 Because the equation expresses the dependence of the variability of returns in future periods from previous periods, we denote this variability as conditional.

3 Requires the *t*-series package. This can be downloaded for free from http://www.r-project.org/

13 STOCHASTIC DIFFERENTIAL EQUATIONS FOR DERIVATIVE PRICING AND ENERGY RISK MANAGEMENT

1 Notice in this case we plot the probabilities associated with the Lognormal distribution versus the price data.

2 See as Merton (1982) for further discussion of this issue.

3 We have already seen a number of approaches to model "fat tails," Using a probability distribution with tails fatter than the normal (see Chapter 5), utilizing distribution with time-varying parameters via the GARCH class of models (see Chapter 12) the application of simple mixture distributions (see Chapter 5) and nonparametric methods (see Chapter 6).

4 Such as a reduction in crude oil output due to political upheaval in the Middle East or generating/transmission constraints in the case of electricity.

5 The probability density function of the Poisson distribution is given by:

$$f(q) = \lambda^q \frac{\exp(-\lambda)}{q!} \quad \text{where } q \geq 0$$

where $1/\lambda$ can be interpreted as the arithmetic mean.

6 It is interesting to note that the density of returns generated from this process at any fixed point in time are in fact mixtures of the exponential family of distributions. Furthermore, since the dynamics of the asset price are driven by two sources of uncertainty, from a mathematical standpoint the model is incomplete.

7 Kat and Heynen (1994) find that the best volatility predictions for equity indices are generated by the stochastic volatility model. For currencies on the other hand, they find that the best forecasts come from the GARCH(1,1) model.

Index