Investment Analysis

Gareth D. Myles

May 2008

Contents

Preface

This book has developed from the lectures for a final-year undergraduate course and a first-level graduate course in finance that I have taught at the University of Exeter for a number of years. They present the essential elements of investment analysis as a practical tool with a firm theoretical foundation. This should make them useful for those who wish to learn investment techniques for practical use and those wishing to progress further into the theory of finance. The book avoids making unnecessary mathematical demands upon the reader but it does treat finance as an analytical tool. The material in the book should be accessible to anyone with undergraduate courses in principles of economics, mathematics and statistics.

Introduction

Finance, and the theory of finance, are important. Why? Because of the growth of financial markets around the world, the volume of trade and the opportunities for profit. Finance theory is about the construction and management of portfolios. This is helped by understanding theories of finance including the pricing of derivatives.

The notes have an emphasis on calculation - of returns, variances etc. They treat finance as an analytical subject but recognize the role and limitation of theory.

Part I

Investment Fundamentals

Chapter 1

Securities and Analysis

Learning investment analysis is a journey into a wealth of knowledge that is an exciting mix of the practical and the analytical. It looks to technique to evaluate and to theory to explain. It is natural to feel a degree of trepidation at the start of such a journey. To help offset this we need to familiarize ourselves with the landscape and landmarks, to develop an overview of our route. Some of these landmarks may be familiar others may be new or be seen from a different perspective. Armed with this we can map out our route.

1.1 Introduction

This book is about the investment of wealth in financial securities. It provides an introduction to the tools of investment analysis that can be used to guide informed investment decisions. These tools range from the knowledge of the securities that are available and how they are traded, through the techniques for evaluating investments, to theories of market functioning.

Some investments can be very successful. An investor placing $10,000 in August 1998 in the stock of Cephalon, a biopharmaceutical company traded on Nasdaq, would have stock worth $107,096 in September 2003. Similarly, a purchase of £10,000 in September 2001 of Lastminute.com stock, an internet retailer traded on the London Stock Exchange, would be worth £134,143 in August 2003. Cephalon and Lastminute.com are far from being alone in offering these levels of gain. Many high technology companies match and can even outstrip their performance. On the down side, losses in value can be even more spectacular. Anyone investing $10,000 in September 2000 in Palm Inc., the makers of handheld computers also traded on Nasdaq, would see that reduced to $91 in April 2003. Such falls are not restricted to manufacturers. A holding in July 2000 of £15 million in Exeter Equity Growth Fund would be worth £72,463 in August 2003 due to a fall in share price from 103.50 to 0.50.

What can be learnt from this book that would help choose investments like

Cephalon and avoid Palm Inc.? The honest answer is that in September 2000 none of the evidence and none of the tools of investment analysis could have forewarned that the stock of Palm Inc. would collapse in the way it did. Far from being a condemnation of the methods, this observation shows precisely why they are so valuable. How can this be so? Because it emphasizes that the world of investments is plagued by uncertainty and unpredictability. No matter how sophisticated are the tools we develop, or how rigorously we do our research into potential investments, it is not possible for an investor to predict the future. And that, in a nutshell, is why we need to learn investment analysis.

Investment analysis encompasses a methodology for accommodating the fundamental uncertainty of the financial world. It provides the tools that an investor can employ to evaluate the implications of their portfolio decisions and gives guidance on the factors that should be taken into account when choosing a portfolio. Investment analysis cannot eliminate the uncertainty, but it can show how to reduce it. Moreover, although it cannot guarantee to guide you to winners like Cephalon, it can help stop you being the investor that places all their wealth in Palm Inc.

The starting point for investment analysis is the market data on the values of securities which describes how they have performed in the past. In some parts of the book, this market data is taken as given and we study how we should invest on the basis of that data. This generates a set of tools which, even if an investor does not apply them literally, provide a powerful framework in which to think rationally about investment. This framework continually emphasizes why many regretful investors have found to their cost that the maxim "there is no such thing as a free lunch" is especially true in financial markets.

A serious investor will want to go beyond just accepting market data and progress to an understanding of the forces that shape the data. This is the role of financial theories that investigate explanations for what is observed. The deeper understanding of the market encouraged by theory can benefit an investor by, at the very least, preventing costly mistakes. The latter is especially true in the world of derivative securities we meet later. But a theory remains just that until it has been shown to unequivocally fit the data, and the wise investor should never forget the limitations of theoretical explanations.

The book will provide information on how to choose which securities to invest in, how they are traded, and the issues involved in constructing and evaluating a portfolio. Throughout the text examples draw on the freely-available and extensive data from Yahoo and show how the methods described can be applied to this data.

1.2 Financial Investment

It is helpful to begin the analysis with a number of definitions that make precise the subject matter that we will be studying. A standard definition is that *investment is the sacrifice of current consumption in order to obtain increased consumption at a later date.* From this perspective, an investment is undertaken with the expectation that it will lead, ultimately, to a preferred pattern of consumption for the investor.

This definition makes consumption the major motivation for investment. In contrast, many investors would argue that their motivation for investment is to increase their wealth. This observation can be related back to the definition by noting that wealth permits consumption or, in more formal language, an increase in the *stock* of wealth permits an increase in the *flow* of consumption. Wealth and consumption are, therefore, two sides of the same coin.

Looking more closely, two different forms of investment can be identified. *Real investment* is the purchase of physical capital such as land and machinery to employ in a production process and earn increased profit. In contrast, *financial investment* is the purchase of "paper" securities such as stocks and bonds.

We do not explicitly discuss real investments in this book. Firms undertake real investment to generate the maximum profit given the market conditions that they face. There are many interesting issues raised by the real investment activities of firms including issues of research and development, capacity expansion, and marketing. But consideration of these matters falls strictly outside the scope of a text whose focus is upon financial investment. It should be noted, though, that a real investment by an individual, such as the purchase of a house or a painting, must be considered as part of the overall portfolio of assets held by that investor.

There are, however, links between the two forms of investment. For example, the purchase of a firm's shares is a financial investment for those who buy them but the motive for the issue of the shares is invariably that the firm wishes to raise funds for real investment. Similarly, the commitment of a householder to a mortgage, which is a financial investment, generates funds for a real investment in property.

As a brief preview, the issues concerning financial investment that are addressed in the following chapters include:

- The forms of security available: where and how they are bought and sold;
- The investment process: the decision about which securities to purchase, and how much of each;
- Financial theory: the factors that determine the rewards from investment and the risks.

The strategy employed to address these issues has the following structure. The first step is to introduce the most important forms of securities that are available to the investor and the ways in which they can be traded. The next step is to analyze the general issues that are involved determining the preferred choice of investment. This is undertaken abstracting from the particular features of different securities. Next, we consider financial theories that try to explain what is observed in the financial markets and which provide further insight into the investment decision. Finally, we return to detailed analysis of some special types of securities that raise especially interesting analytical questions.

1.3 Investment Analysis

The purpose of this book is to teach the principles of investment analysis. So, what is investment analysis? One definition that moves us a little way forward is that:

> *"Investment analysis is the study of financial securities for the purpose of successful investing."*

This definition contains within it a number of important points. Firstly, there are the institutional facts about financial securities: how to trade and what assets there are to trade. Secondly, there are analytical issues involved in studying these securities: the calculation of risks and returns, and the relationship between the two. Then there is the question of what success means for an investor, and the investment strategies that ensure the choices made are successful. Finally, there are the financial theories that are necessary to try to understand how the markets work and how the prices of assets are determined.

It is clear that the more an investor understands, the less likely they are to make an expensive mistake. Note carefully that this is not saying that the more you know, the more you will earn. An explanation for this observation will be found in some of the theories that follow. These comments partly address the question "Can you beat the market?" Whether you can depends on the view you may hold about the functioning of financial markets. One of the interpretations of investment analysis is that this is just not possible on a repeated basis. An alternative interpretation is that knowing the theory reveals where we should look for ways of beating the market.

Example 1 *The website for GinsGlobal Index Funds puts it this way "Very few professional fund managers can beat the market. Since there is no reliable way to identify the fund managers who will outperform the market, investors are best served by buying a broad spectrum of stocks at lower cost" (www.ginsglobal.co.za/company_profile.htm).*

A knowledge of investment analysis can be valuable in two different ways. It can be beneficial from a personal level. The modern economy is characterized by ever increasing financial complexity and extension of the range of available securities. Moreover, personal wealth is increasing, leading to more funds that private individuals must invest. There is also a continuing trend towards greater reliance on individual provision for retirement. The wealth required for retirement must be accumulated whilst working and be efficiently invested.

The study of investment analysis can also provide an entry into a rewarding professional career. There are many different roles for which investment analysis is useful and the material covered in this book will be useful for many of them. The training to become a financial analyst requires knowledge of much of this analysis. Further, there are positions for brokers, bankers and investment advisors for whom knowledge of investment analysis is a distinct advantage.

Example 2 *The Association for Investment Management and Research (AIMR) is an international organization of over 50,000 investment practitioners and educators in more than 100 countries. It was founded in 1990 from the merger of the Financial Analysts Federation and the Institute of Chartered Financial Analysts. It oversees the Chartered Financial Analyst (CFA®) Program which is a globally-recognized standard for measuring the competence and integrity of financial analysts. CFA exams are administered annually in more than 70 countries. (For more information, see www.aimr.org)*

1.4 Securities

A security can be defined as:

> *"A legal contract representing the right to receive future benefits under a stated set of conditions."*

The piece of paper (*e.g.* the share certificate or the bond) defining the property rights is the physical form of the security. The terms *security* or *asset* can be used interchangeably. If a distinction is sought between them, it is that the term assets can be applied to both financial and real investments whereas a security is simply a financial asset. For much of the analysis it is asset that is used as the generic term.

From an investor's perspective, the two most crucial characteristics of a security are the *return* it promises and the *risk* inherent in the return. An informal description of return is that it is the gain made from an investment and of risk that it is the variability in the return. More precise definitions of these terms and the methods for calculating them are discussed in Chapter 3. For the present purpose, the return can be defined as the percentage increase in the value of the investment, so

$$\text{Return} = \frac{\text{final value of investment} \; - \; \text{initial value of investment}}{\text{initial value of investment}} \times 100. \quad (1.1)$$

Example 3 *At the start of 2003 an investor purchased securities worth \$20000. These securities were worth \$25000 at the end of the year. The return on this investment is*

$$\textit{Return} = \frac{\textit{25000 - 20000}}{\textit{20000}} \times 100 = 25\%.$$

The return on a security is the fundamental reason for wishing to hold it. The return is determined by the payments made during the lifetime of the security plus the increase in the security's value. The importance of risk comes from the fact that the return on most securities (if not all) is not known with certainty when the security is purchased. This is because the future value of security is unknown and its flow of payments may not be certain. The risk of a security is a measure of the size of the variability or uncertainty of its return.

It is a fundamental assumption of investment analysis that investors wish to have more return but do not like risk. Therefore to be encouraged to invest in assets with higher risks they must be compensated with greater return. This fact, that increased return and increased risk go together, is one of the fundamental features of assets.

A further important feature of a security is its *liquidity*. This is the ease with which it can be traded and turned into cash. For some assets there are highly developed markets with considerable volumes of trade. These assets will be highly liquid. Other assets are more specialized and may require some effort to be made to match buyers and sellers. All other things being equal, an investor will always prefer greater liquidity in their assets.

The major forms of security are now described. Some of these are analyzed in considerably more detail in later chapters because they raise interesting questions in investment analysis.

1.5 Non-Marketable Securities

The first form of security to introduce are those which are non-marketable, meaning that they cannot be traded once purchased. Despite not being tradeable, they are important because they can compose significant parts of many investors' portfolios.

The important characteristics of these securities are that they are personal - the investor needs to reveal personal details in order to obtain them so that the parties on both sides know who is involved. They tend to be safe because they are usually held at institutions that are insured and are also liquid although sometimes at a cost.

The first such security is the *savings account*. This is the standard form of deposit account which pays interest and can be held at a range of institutions from commercial banks through to credit unions. The interest rate is typically variable over time. In addition, higher interest will be paid as the size of deposit increases and as the notice required for withdrawal increases. Withdrawals can sometimes be made within the notice period but will be subject to penalties.

A second significant class are *government savings bonds*. These are the non-traded debt of governments. In the US these are purchased from the Treasury indirectly through a bank or savings institution. The bonds receive interest only when they are redeemed. Redemption is anytime from six months after the issue date. National Savings in the UK deal directly with the public and offers a variety of bonds with different returns, including bonds with returns linked to a stock exchange index.

Two other securities are *non-negotiable certificates of deposit* (CDs). These are certificates issued by a bank, savings and loan association, credit union, or similar financial organization that confirm that a sum of money has been received by the issuer with an implied agreement that the issuer will repay the sum of money and that they are not a negotiable (or tradeable) instrument. CDs can have a variety of maturities and penalties for withdrawal. They are essentially a loan from an investor to a bank with interest paid as the reward. A *money market deposit account* (MMDA) is an interest-earning savings account offered by an insured financial institution with a minimum balance requirement. The special feature of the account is that is has limited transaction privileges: the investor is limited to six transfers or withdrawals per month with no more than three transactions as checks written against the account. The interest rate paid on a MMDA is usually higher than the rate for standard savings account.

1.6 Marketable Securities

Marketable securities are those that can be traded between investors. Some are traded on highly developed and regulated markets while others can be traded between individual investors with brokers acting as middle-men.

This class of securities will be described under four headings. They are classified into *money market securities* which have short maturities and *capital market securities* which have long maturities. The third group are *derivatives* whose values are determined by the values of other assets. The final group are classified as *indirect investments* and represent the purchase of assets via an investment company.

1.6.1 Money Market Securities

Money market securities are short-term debt instruments sold by governments, financial institutions and corporations. The important characteristic of these securities is that they have maturities when issued of one year or less. The minimum size of transactions is typically large, usually exceeding $100,000.

Money market securities tend to be highly liquid and safe assets. Because of the minimum size of transactions, the market is dominated by financial institutions rather than private investors. One route for investors to access this market is via money market mutual funds.

Treasury Bills

Short-term treasury bills are sold by most governments as a way of obtaining revenues and for influencing the market. As later chapters will show, all interest rates are related so increasing the supply of treasury bills will raise interest rates (investors have to be given a better reward to be induced to increase demand) while reducing the supply will lower them.

Treasury bills issued by the US federal government are considered to be the least risky and the most marketable of all money markets instruments. They represent a short-term loan to the US federal government. The US federal government has no record of default on such loans and, since it can always print money to repay the loans, is unlikely to default. Treasury bills with 3-month and 6-month maturities are sold in weekly auctions. Those with a maturity of 1 year are sold monthly. Treasury Bills have a face value of $1000 which is the amount paid to the holder at the maturity date. They sell at a *discount* (meaning a price different to, and usually less, than face value) and pay no explicit interest payments. The benefit to the investor of holding the bill is the difference between the price paid and the face value received at maturity.

An important component in some of the analysis in the later chapters is the *risk-free asset*. This is defined as an asset which has a known return and no risk. Because US Treasury Bills (and those of other governments with a similar default-free record) are considered to have no risk of default and a known return, they are the closest approximations that exist to the concept of a risk-free investment. For that reason, the return on Treasury Bills is taken as an approximation of the risk-free rate of return.

Commercial Paper

Commercial paper is a short term promissory note issued by a corporation, typically for financing accounts for which payment is due to be received and for financing inventories. The value is usually at least $100,000 and the maturity 270 days or less. They are usually sold at a discount. These notes are rated by ratings agencies who report on the likelihood of default.

Eurodollars

Eurodollars are dollar-denominated deposits held in non-US banks or in branches of US banks located outside the US. Because they are located outside the US, Eurodollars avoid regulation by the Federal Reserve Board. Eurodollars originated in Europe but the term also encompasses deposits in the Caribbean and Asia. Both time deposits and CDs can fall under the heading of Eurodollars. The maturities are mostly short term and the market is mainly between financial institutions. The freedom from regulation allows banks in the Eurodollar market to operate on narrower margins than banks in the US. The market has expanded as a way of avoiding the regulatory costs of dollar-denominated financial intermediation.

Negotiable Certificates of Deposit

As for non-negotiable CDs, these are promissory notes on a bank issued in exchange for a deposit held in a bank until maturity. They entitle the bearer to receive interest. A CD bears a maturity date (mostly 14 days to 1 year), a specified interest rate, and can be issued in any denomination. CDs are generally issued by commercial banks. These CDs are tradeable with dealers *making a market* (meaning they buy and sell to give the market liquidity). CDs under $100,000 are called "small CDs," CDs for more than $100,000 are called "large CDs" or "Jumbo CDs."

Bankers Acceptance

A bankers acceptance is a short-term credit investment created by a non-financial firm but which is guaranteed by a bank. The acceptances can be traded at discounts from face value. Maturities range from 30 - 180 days and the minimum denomination is $100,000. Bankers' Acceptance are very similar to treasury bills and are often used in money market funds.

Repurchase Agreements

A repurchase agreement involves a dealer selling government securities to an investor with a commitment to buy them back at an agreed time. The maturity is often very short with many repurchase agreement being overnight. They constitute a form of short term borrowing for dealers in government securities. The interest rate on the transaction is the difference between the selling and repurchase prices. They permit the dealer to attain a *short position* (a negative holding) in bonds.

1.6.2 Capital Market Securities

Capital market securities include instruments having maturities greater than one year and those having no designated maturity at all. In the latter category can be included common stock and, in the UK, consuls which pay a coupon in perpetuity. The discussion of capital market securities divides them into fixed income securities and equities.

Fixed Income Securities

Fixed income securities promise a payment schedule with specific dates for the payment of interest and the repayment of principal. Any failure to conform to the payment schedule puts the security into default with all remaining payments. The holders of the securities can put the defaulter into bankruptcy.

Fixed income securities differ in their promised returns because of differences involving the maturity of the bonds, the callability, the creditworthiness of the issuer and the taxable status of the bond. Callability refers to the possibility that the issuer of the security can call it in, that is pay off the principal prior to maturity. If a security is callable, it will have a lower price since the issuer will only call when it is in their advantage to do so (and hence against the interests of the holder). Creditworthiness refers to the predicted ability of the issuer to meet the payments. Income and capital gains are taxed differently in many countries, and securities are designed to exploit these differences. Also, some securities may be exempt from tax.

Bonds Bonds are fixed income securities. Payments will be made at specified time intervals unless the issuer defaults. However, if an investor sells a bond before maturity the price that will be received is uncertain.

The par or face value is usually \$1000 in the US and £100 in the UK. Almost all bonds have a term - the maturity date at which they will be redeemed.

Coupon bonds pay periodic interest. The standard situation is for payment every 6 months. Zero coupon or discount bonds pay no coupon but receive the par value at maturity. The return on a discount bond is determined by the difference between the purchase price and the face value. When the return is positive, the purchase price must be below the face value. Hence, these bonds are said to sell at a discount.

Bonds sell on accrued interest basis so the purchaser pays the price plus the interest accrued up until the date of purchase. If this was not done, sales would either take place only directly after coupon payments or else prices would be subject to downward jumps as payment dates were passed.

Treasury Notes and Bonds The US government issues fixed income securities over a broad range of the maturity spectrum through the Treasury. These are considered safe with no practical risk of default. Treasury notes have a term of more than one year, but no more than 10 years. Treasury bonds have maturities that generally lie in the range of 10 - 30 years.

Notes and bonds are sold at competitive auctions. They sell at face value with bids based on returns. Both notes and bonds pay interest twice a year and repay principal on the maturity date.

Similar notes and bonds are issued by most governments. In the UK, government bonds are also known as gilts since the original issues were gilt-edged. They are sold both by tender and by auction.

Federal Agency Securities Some federal agencies are permitted to issue debt in order to raise funds. The funds are then used to provide loans to assist specified sectors of the economy. The are two types of such agencies: federal agencies and federally-sponsored agencies.

Federal agencies are legally part of the federal government and the securities are guaranteed by the Treasury. One significant example is the National Mortgage Association.

Federally-sponsored agencies are privately owned. They can draw upon the Treasury up to an agreed amount but the securities are not guaranteed. Examples are the Farm Credit System and the Student Loan Marketing Association.

Municipal Bonds A variety of political entities such as states, counties, cities, airport authorities and school districts raise funds to finance projects through the issue of debt. The credit ratings of this debt vary from very good to very poor. Two types of bonds are provided. General obligation bonds are backed by the "full faith and credit" whereas revenue bonds are financed through the revenue from a project.

A distinguishing feature of these bonds is that they are exempt from federal taxes and usually exempt from the taxes of the state issuing the bond.

Corporate Bonds Corporate bonds are similar to treasury bonds in their payment patterns so they usually pay interest at twice yearly intervals. The major difference form government bonds is that corporate bonds are issued by business entities and thus have a higher risk of default. This leads them to be rated by rating agencies.

Corporate bonds are senior securities which means that they have priority over stocks in the event of bankruptcy. Secured bonds are backed by claims on specific collateral but unsecured are backed only by the financial soundness of the corporation. Convertible bonds can be converted to shares when the holder chooses.

Common Stock (Equity)

Common stock represents an ownership claim on the earnings and assets of a corporation. After holders of debt claims are paid, the management of the company can either pay out the remaining earnings to stockholder in the form of dividends or reinvest part or all of the earnings. The holder of a common stock has limited liability. That is, they are not responsible for any of the debts of a failed firm.

There are two main types of stocks: common stock and preferred stock. The majority of stock issued is common stock which represent a share of the ownership of a company and a claim on a portion of profits. This claim is paid in the form of dividends. Stockholders receive one vote per share owned in elections to the company board. If a company goes into liquidation, common stockholders do not receive any payment until the creditors, bondholders, and preferred shareholders are paid.

Preferred Stock

Preferred stock also represents a degree of ownership but usually doesn't carry the same voting rights. The distinction to common stock is that preferred stock has a fixed dividend and, in the event of liquidation, preferred shareholders are paid before the common shareholder. However, they are still secondary to debt holders. Preferred stock can also be callable, so that a company has the option of purchasing the shares from shareholders at anytime. In many ways, preferred stock fall between common stock and bonds.

1.6.3 Derivatives

Derivatives are securities whose value derives from the value of an underlying security or a basket of securities. They are also known as *contingent claims*, since their values are contingent on the performance of the underlying assets.

Options

An *option* is a security that gives the holder the right to either buy (a call option) or sell (a put option) a particular asset at a future date or during a particular period of time for a specified price - *if* they wish to conduct the transactions. If the option is not exercised within the time period then it expires.

Futures

A *future* is the obligation to buy or sell a particular security or bundle of securities at a particular time for a stated price. A future is simply a delayed purchase or sale of a security. Futures were originally traded for commodities but now cover a range of financial instruments.

Rights and Warrants

Contingent claims can also be issued by corporations. Corporate-issued contingent claims include *rights* and *warrants*, which allow the holder to purchase common stocks from the corporation at a set price for a particular period of time.

Rights are securities that give stockholders the entitlement to purchase new shares issued by the corporation at a predetermined price, which is normally less than the current market price, in proportion to the number of shares already owned. Rights can be exercised only within a short time interval, after which they expire.

A *warrant* gives the holder the right to purchase securities (usually equity) from the issuer at a specific price within a certain time interval. The main distinction between a warrant and a call options is that warrants are issued and guaranteed by the corporation, whereas options are exchange instruments. In addition, the lifetime of a warrant can be much longer than that of an option.

1.6.4 Indirect Investments

Indirect investing can be undertaken by purchasing the shares of an *investment company*. An investment company sells shares in itself to raise funds to purchase a portfolio of securities. The motivation for doing this is that the pooling of funds allows advantage to be taken of diversification and of savings in transaction costs. Many investment companies operate in line with a stated policy objective, for example on the types of securities that will be purchased and the nature of the fund management.

Unit Trusts

A *unit trust* is a registered trust in which investors purchase units. A portfolio of assets is chosen, often fixed-income securities, and passively managed by a professional manager. The size is determined by inflow of funds. Unit trusts are designed to be held for long periods with the retention of capital value a major objective.

Investment Trusts

The *closed-end investment trust* issue a certain fixed sum of stock to raise capital. After the initial offering no additional shares are sold. This fixed capital is then managed by the trust. The initial investors purchase shares, which are then traded on the stock market.

An *open-end investment company* (or *mutual fund*) continues to sell shares after the initial public offering. As investors enter and leave the company, its capitalization will continually change. Money-market funds hold money-market instrument while stock and bond and income funds hold longer-maturity assets.

Hedge Funds

A *hedge fund* is an aggressively managed portfolio which takes positions on both safe and speculative opportunities. Most hedge funds are limited to a maximum of 100 investors with deposits usually in excess of $100,000. They trade in all financial markets, including the derivatives market.

1.7 Securities and Risk

The risk inherent in holding a security has been described as a measure of the size of the variability, or the uncertainty, of its return. Several factors can be isolated as affecting the riskiness of a security and these are now related to the securities introduced above. The comments made are generally true, but there will always be exceptions to the relationships described.

- *Maturity* The longer the period until the maturity of a security the more risky it is. This is because underlying factors have more chance to change over a longer horizon. The maturity value of the security may be eroded by inflation or, if it is denominated in a foreign currency, by currency fluctuations. There is also an increased chance of the issuer defaulting the longer is the time horizon.

- *Creditworthiness* The governments of the US, UK and other developed countries are all judged as safe since they have no history of default in the payment of their liabilities. Therefore they have the highest levels of creditworthiness being judged as certain to meet their payments schedules. Some other countries have not had such good credit histories. Both Russia and several South American countries have defaulted in the recent past. Corporations vary even more in their creditworthiness. Some are so lacking in creditworthiness that an active "junk bond" market exists for high return, high risk corporate bonds that are judged very likely to default.

- *Priority* Bond holders have the first claim on the assets of a liquidated firm. Only after bond holders and other creditors have been paid will stock holders receive any residual. Bond holders are also able to put the corporation into bankruptcy if it defaults on payment. This priority reduces the risk on bonds but raises it for common stock.

- *Liquidity* Liquidity relates to how easy it is to sell an asset. The existence of a highly developed and active secondary market raises liquidity. A security's risk is raised if it is lacking liquidity.

- *Underlying Activities* The economic activities of the issuer of the security can affect its riskiness. For example, stock in small firms and in firms operating in high-technology sectors are on average more risky than those of large firms in traditional sectors.

These factors can now be used to provide a general categorization of securities into different risk classes.

Treasury bills have little risk since they represent a short-term loan to the government. The return is fixed and there is little chance of change in other prices. There is also an active secondary market. Long-term government bonds have a greater degree of risk than short-term bonds. Although with US and UK government bonds there is no risk of default and the percentage payoff is fixed, there still remains some risk. This risk is due to inflation which causes uncertainty in the real value of the payments from the bond even though the nominal payments are certain.

The bonds of some other countries bonds may have a risk of default. Indeed, there are countries for which this can be quite significant. As well as an inflation risk, holding bonds denominated in the currency of another country leads to an exchange rate risk. The payments are fixed in the foreign currency but this does not guarantee their value in the domestic currency. Corporate bonds suffer from inflation risk as well as an enhanced default risk relative to government bonds.

Common stocks generally have a higher degree of risk than bonds. A stock is a commitment to pay periodically a dividend, the level of which is chosen by the firm's board. Consequently, there is no guarantee of the level of dividends. The risk in holding stock comes from the variability of the dividend and from the variability of price.

Generally, the greater the risk of a security, the higher is expected return. This occurs because return is the compensation that has to be paid to induce investors to accept risks. Success in investing is about balancing risk and return to achieve an optimal combination.

1.8 The Investment Process

The investment process is description of the steps that an investor should take to construct and manage their portfolio. These proceed from the initial task of identifying investment objectives through to the continuing revision of the portfolio in order to best attain those objectives.

The steps in this process are:

1. Determine Objectives. Investment policy has to be guided by a set of objectives. Before investment can be undertaken, a clear idea of the purpose of the investment must be obtained. The purpose will vary between investors. Some may be concerned only with preserving their current wealth. Others may see investment as a means of enhancing wealth. What primarily drives objectives is the attitude towards taking on risk. Some investors may wish to eliminate risk as much as is possible, while others may be focussed almost entirely on return and be willing to accept significant risks.

2. Choose Value The second decision concerns the amount to be invested. This decision can be considered a separate one or it can be subsumed in the allocation decision between assets (what is not invested must either be held in some other form which, by definition, is an investment in its own right or else it must be consumed).

3. Conduct Security Analysis. Security analysis is the study of the returns and risks of securities. This is undertaken to determine in which classes of assets investments will be placed and to determine which particular securities should be purchased within a class. Many investors find it simpler to remain with the more basic assets such as stocks and fixed income securities rather than venture into complex instruments such as derivatives. Once the class of assets has been determined, the next step is to analyze the chosen set of securities to identify relevant characteristics of the assets such as their expected returns and risks. This information will be required for any informed attempt at portfolio construction.

Another reason for analyzing securities is to attempt to find those that are currently mispriced. For example, a security that is under-priced for the returns it seems to offer is an attractive asset to purchase. Similarly, one that is over-priced should be sold. Whether there are any assets are underpriced depends on the degree of efficiency of the market. More is said on this issue later.

Such analysis can be undertaken using two alternative approaches:

- *Technical analysis* This is the examination of past prices for predictable trends. Technical analysis employs a variety of methods in an attempt to find patterns of price behavior that repeat through time. If there is such repetition (and this is a disputed issue), then the most beneficial times to buy or sell can be identified.

- *Fundamental analysis* The basis of fundamental analysis is that the true value of a security has to be based on the future returns it will yield. The analysis allows for temporary movements away from this relationship but requires it to hold in the long-rum. Fundamental analysts study the details of company activities to makes predictions of future profitability since this determines dividends and hence returns.

4. Portfolio Construction. Portfolio construction follows from security analysis. It is the determination of the precise quantity to purchase of each of the chosen securities. A factor that is important to consider is the extent of *diversification*. Diversifying a portfolio across many assets may reduce risk but it involves increased transactions costs and increases the effort required to manage the portfolio. The issues in portfolio construction are extensively discussed in Chapters 4 and 5.

5. Evaluation. Portfolio evaluation involves the assessment of the performance of the chosen portfolio. To do this it is necessary to have some yardstick for comparison since a meaningful comparison is only achieved by comparing the return on the portfolio with that on other portfolios with similar risk characteristics. Portfolio evaluation is discussed in Chapter 17.

6. Revision. Portfolio revision involves the application of all the previous steps. Objectives may change, as may the level of funds available for investment. Further analysis of assets may alter the assessment of risks and returns and new assets may become available. Portfolio revision is therefore the continuing re-application of the steps in the investment process.

1.9 Summary

This chapter has introduced investment analysis and defined the concept of a security. It has looked at the securities that are traded and where they are traded. In addition, it has begun the development of the concepts of risk and return that characterize securities. The fact that these are related - an investor cannot have more of one without more of another - has been stressed. This theme will recur throughout the book. The chapter has also emphasized the role of uncertainty in investment analysis. This, too, is a continuing theme.

It is hoped that this discussion has provided a convincing argument for the study of investment analysis. Very few subjects combine the practical value of investment analysis with its intellectual and analytical content. It can provide a gateway to a rewarding career and to personal financial success.

Exercise 1 *Use the monthly data on historical prices in Yahoo to confirm the information given on the four stocks in the Introduction. Can you find a stock that has grown even faster than Cephalon?*

Exercise 2 *There are many stocks which have performed even worse than Exeter Equity Growth Fund. Why will many of these be absent from the Yahoo data?*

Exercise 3 *If a method was developed to predict future stock prices perfectly, what effect would it have upon the market?*

Exercise 4 *At the start of January 1999 one investor makes a real investment by purchasing a house for $300000 while a second investor purchases a portfolio of securities for $300000. The first investor lives in the house for the next two years. At the start of January 2001 the house is worth $350000 and the portfolio of securities is worth $375000. Which investor has fared better?*

Exercise 5 *Is a theory which tells us that we "cannot beat the market" useless?*

Exercise 6 *You are working as a financial advisor. A couple close to retirement seek your advice. Should you recommend a portfolio focused on high-technology stock or one focused on corporate bonds? Would your answer be different if you were advising a young newly-wed couple?*

Exercise 7 *Obtain a share certificate and describe the information written upon it.*

Exercise 8 *By consulting the financial press, obtain data on the interest rates on savings accounts. How are these rates related to liquidity?*

Exercise 9 *Taking data on dividends from Yahoo, assess whether the prices of stocks are related to their past dividend payments. What does your answer say about fundamental analysis?*

Exercise 10 *If all investors employed technical analysis, would technical analysis work?*

Exercise 11 *Are US treasury bills a safe asset for an investor who lives in Argentina?*

Exercise 12 *Corporations usually try to keep dividend payments relatively constant even in periods when profits are fluctuating. Why should they wish to do this?*

Chapter 2

Buying and Selling

In chess, after learning the names of the pieces, the next step is to understand the moves that the pieces may make. The ability of each piece to move in several ways provides the complexity of the game that has generated centuries of fascination. By combining these moves, chess manuals describe the standard openings, the philosophies of the middle game and the killer finishes. Similar rules apply to trading securities. Much more is involved than simply buying and selling. Getting to know the rules of the game and the trades that can be made will help the investor just as much as it helps the chess player.

2.1 Introduction

A fundamental step in the investment process is the purchase and sale of securities. There is more to this than is apparent at first sight. An order to buy or sell can take several forms, with characteristics that need to be determined by the investor. A variety of brokers with different levels of service, and corresponding fees, compete to act on the investor's behalf. Some brokers are even prepared to loan funds for the investor to purchase assets.

The chapter begins with a discussion of the markets on which securities are traded. The role and characteristics of brokers are then described. Following this, the focus turns to the purchase of common stock since it is here that there is the greatest variety of purchasing methods. The choice of method can affect the return on a portfolio just as significantly as can the choice of asset so the implications for returns are considered.

2.2 Markets

Securities are traded on markets. A market is a place where buyers and sellers of securities meet or any organized system for connecting buyers and sellers.

Markets are fundamental for the trading of securities.

Markets can have a physical location such as the New York Stock Exchange or the London International Financial Futures Exchange. Both of these have a trading floor where trade is conducted. It is not necessary for there to be a physical location. The London Stock Exchange once possessed a physical location, but now trade is conducted through a computer network that links dealers. The Nasdaq Stock Market also has no location but relies on a network to link dealers. Recent innovations such as internet-based markets also have no physical location.

Example 4 *The New York Stock Exchange was founded in 1792 and registered as a national securities exchange with the U.S. Securities and Exchange Commission on October 1, 1934. It was incorporated as a not-for-profit corporation in 1971. The Exchange building at 18 Broad Street was opened in 1903 and a number of additional buildings are now also in use. At the end of 2002, 2,959 stocks were listed with a combined value of $9,603.3 billion. In July 2003, 31,924.5 million shares were traded with a combined value of $896.0 billion and an average share price of $28.07. Only members of the Exchange can trade and to become a member a "seat" must be purchased. The highest price paid for an NYSE seat was $2,650,000 on August 23, 1999. (www.nyse.com)*

Example 5 *Nasdaq opened in 1971 as the first electronic market and is currently the largest. It lists just under 4000 companies primarily in the technology, retail, communication, financial services and biotechnology sectors. Information on market activity is transmitted to 1.3 million users in 83 countries. There is an average of 19 market makers for each listed company with Dell Computer Corporation having 95 market makers. Annual share volume in 2002 was 441 billion shares with a value of $7.3 trillion. (www.nasdaq.com)*

Markets can be classified in a number of different ways. Each classification draws out some important aspects of the role and functioning of markets.

2.2.1 Primary and Secondary

Primary markets are security markets where new issues of securities are traded. When a company first offers shares to the market it is called an *initial public offering*. If additional shares are introduced later, they are also traded on the primary market. The price of shares is normally determined through trade but with new shares there is no existing price to observe. The price for initial public offerings has either to be set as part of the offer, or determined through selling the shares by tender or auction.

Secondary markets are markets where existing securities are resold. The London and New York stock exchanges are both primarily secondary markets.

The role of the primary market in helping to attain economic efficiency is clear: the primary market channels funds to those needing finance to undertake real investment. In contrast, the role of the secondary market, and the reason why so much attention is paid to it, is probably less clear. Two important roles for the secondary market that can be identified:

- *Liquidity* One of the aspects that will be important for the purchaser of a new security is their ability to sell it at a later date. If it cannot be sold, then the purchaser is making a commitment for the lifetime of the asset. Clearly, given two otherwise identical assets an investor would prefer to own the one which can most easily be traded. Thus new securities would have a lower value if they could not be subsequently traded. The existence of a secondary market allows such trading and increases the liquidity and value of an asset.

- *Value* Trading in assets reveals information and provides a valuation of those assets. The assignment of values guides investment decisions by showing the most valuable uses for resources and helps in the attainment of economic efficiency. Without the secondary market this information would not be transmitted.

2.2.2 Call and Continuous

A second way to classify markets is by the nature of trading and the time periods at which trading can take place.

In a *call market* trading takes place at specified times. Those who wish to trade are called together at a specific time and trade for a limited period. A single price is set that ensures the market clears. This can cause significant movements in price from one trading time to the next, so call markets can have provisions to limit movement from the initial price.

Example 6 *The main Austrian exchange, Wiener Börse, operates a call system to auction shares. The auction price is set to ensure that the largest volume of orders can be executed leaving as few as possible unfilled. An auction schedule is published to announce the times when specific securities are called. (www.wienerboerse.at)*

In a *continuous market* there is trading at all times the market is open. Requests to buy and sell are made continuously. Trade is often facilitated by *market makers* who set prices and hold inventories.

Example 7 *The London Stock Exchange operates as a continuous market and is the largest equity market in Europe. On the London Stock Exchange trading is performed via computer and telephone using dealing rooms that are physically separated from the exchange. Almost 300 firms worldwide trade as members of the Exchange. (www.londonstockexchange.com)*

2.2.3 Auction and Over-the-Counter

In an *auction* market buyers and sellers enter a bidding process to determine the trading price of securities. This typically takes place at a specified location. The New York Stock Exchange is the primary example of an auction market.

An *over-the-counter* market involves direct negotiation between broker and dealers over a computer network or by telephone. The market will have a network of dealers who make a market and are willing to buy and sell at specified prices. They earn profit through the *spread*: the difference between the price at which they will buy and the price at which they will sell (the latter being higher). Nasdaq is considered to be an over-the-counter market.

2.2.4 Money and Capital

The *money market* is the market for assets with a life of less than 1 year. This includes money itself and near-money assets such as short term bonds.

The *capital market* is the market for assets with a life greater than 1 year such as equity and long-term bonds.

2.3 Brokers

On most markets, such as the New York and London Stock Exchanges, an individual investor cannot trade on the market directly. Instead they must employ the services of a broker who will conduct the trade on their behalf. This section discusses brokers and the services offered by brokerages.

A *broker* is a representative appointed by an individual investor to make transactions on their behalf. The reward for a broker is generated through commission charged on the transactions conducted. This can lead to incentive problems since it encourages the broker to recommend excessive portfolio revision or *churning*. The accounts of individual investors at a brokerage are dealt with by an account executive. Institutional investors deal through special sections of retail brokerage firms

Brokerage firms can be classified according to the services offered and the resulting level of fee charged. Traditional brokerages, now called *full-service* brokers, offer a range of services including information, investment advice and investment publications. They conduct the trading business of the clients and aim to guide them with their investment decisions. In addition to earning income from commissions, full-service brokers also generate revenue from a range of other activities. Amongst these are trading on their own account, commission from the selling of investment instruments such as mutual funds and payment for participation in initial public offerings.

Example 8 *In 2002, the assets of the retail customers of Morgan Stanley amounted to $517 billion and they employed 12,500 financial advisors. Their retail brokerage business now focuses on fee-based accounts rather than commission and has changed the incentive structure for financial advisors so that the interests of the investor and the financial advisor coincide. The financial advisors also take a more consultative approach with investors and emphasize financial planning, asset allocation and diversification. Managed investment products such as mutual funds, managed accounts and variable annuities have become a major focus. (www.morganstanley.com)*

Discount brokers offer fewer services and charge lower fees than full-service brokers. Effectively, they do not provide advice or guidance or produce publications. Their major concentration is upon the execution of trading orders. Many discount brokers operate primarily internet-based services.

Example 9 *Quick & Reilly charge a minimum commission rate of $19.95 for orders placed online for stocks priced over $2.00. A higher rate applies to stock priced under $2.00 and for trades executed over the telephone or through financial consultants. A full schedule of fees can be found at www.quickandreilly.com.*

2.4 Trading Stocks

To trade stocks through a broker it is necessary to provide a range of information. Some of this information is obvious, others parts require explanation. The details of the transaction that need to be given to the broker are:

- The name of the firm whose stock is to be traded;
- Whether it is a buy or a sell order;
- The size of the order;
- The time limit until the order is cancelled;
- The type of order.

Of these five items, the first three are self-explanatory. The final two are now explored in more detail.

2.4.1 Time Limit

The time limit is the time within which the broker should attempt to fill the order. Most orders can be filled immediately but for some stocks, such as those for small firms, there may not be a very active market. Also, at times when the market is falling very quickly it may not be possible to sell. In the latter case a time limit is especially important since the price achieved when the order is filled may be very different to when the order was placed.

A *day order* is the standard order that a broker will assume unless it is specified otherwise. When a day order is placed the broker will attempt to fill it during the day that it is entered. If it is not filled on that day, which is very unlikely for an order concerning a sale or purchase of stock in a large corporation, the order is cancelled.

An open-ended time horizon can be achieved by placing an *open order,* also known as a good-till-cancelled order. Such an order remains in effect until it is either filled or cancelled. In contrast, a *fill-or-kill* order is either executed immediately or, if this cannot be done, cancelled. Finally, a *discriminatory order* leaves it to the broker's discretion to decide when to execute or cancel.

2.4.2 Type of Order

The alternative types of order are designed to reduce the uncertainty associated with variations in price.

A *market order* is the simplest transaction. It is a request for the broker to either buy or sell, with the broker making their best effort to complete the transaction and obtain a beneficial price. With a market order the price at which the trade takes place is uncertain but, unless it is for a very illiquid asset, it is usually certain that the broker will complete the transaction.

In a *limit order* a limit price is specified. For a stock purchase, the limit price is the maximum price at which the investor is willing to buy. For a stock sale, the limit price is the minimum they are willing to accept. Execution of a limit order is uncertain since the limit price may be unobtainable. If the transactions does proceed then the upper limit on price (if buying) or the lower limit on price (if selling) is certain.

With a *stop order*, a stop price has to be specified. This *stop price* acts a trigger for the broker to initiate the trade. For a sale, the stop price is set below the market price and the broker is instructed to sell if the price falls below the stop price. A stop-loss strategy of this form is used to lock-in profits. Alternatively, for a buy order, the broker is instructed to buy if the price rises above the stop price (which is set above the current market price). This strategy could be employed by an investor waiting for the best moment to purchase a stock. When its price shows upward movement they then purchase.

The execution of a stop order is certain if the stop price is passed. However the price obtained is uncertain, especially so if there are rapid upward or downward movements in prices.

A *stop-limit order* combines the limit order and the stop order. A minimum price is placed below the stop price for a sell and a maximum price is placed above the stop-price for a buy. This has the effect of restricting price to be certain within a range but execution is uncertain since no transaction may be possible within the specified range.

2.5 Accounts

Before common stock can be through a broker it is first necessary to open an account with a brokerage. This can be done by either physically visiting the brokerage, by telephone or directly by the internet. It is necessary that some personal details are given to the broker.

Example 10 *The online account application form at Quick and Reilly requires answers to five categories of question. These are: (i) personal details including citizenship and social security number; (ii) financial details including income, source of funds and investment objectives; (iii) details of current broker; (iv) employment status; and (v) links with company directors and stock exchange members.*

2.5.1 Account Types

When opening an account at a brokerage, an investor has a choice between the two types of account. A *cash account* requires that the investor provides the entire funds for any stock purchase. In contrast, a *margin account* with a broker allows the investor to borrow from the broker to finance the purchase of assets. This allows a portfolio to be partly financed by using borrowed funds. The implications of this will be analyzed after first considering some further details of margin accounts.

To open a margin account a *hypothecation agreement* is required. Under such an agreement the investor has to agree that the brokerage can:

- Pledge securities purchased using the margin account as collateral;
- Lend the purchased securities to others.

To make this possible, the shares are held in *street name* by the brokerage. This means that they are owned legally by brokerage but dividends, voting rights and annual reports of the companies whose stock are purchased go to the investor. In consequence, the investor receives all the privileges of owning the stock even if they do not legally own it.

The reason that the shares can be pledged as collateral is because the brokerage requires some security for the loan it has advanced the investor. There is always a possibility that an investor may default on the loan, so the brokerage retains the stock as security. Allowing the shares to be lent to other investors may seem a strange requirement. However, this is necessary to permit the process of short-selling to function. This is discussed in Section 2.6.

A margin purchase involves the investor borrowing money from the broker to invest. The broker charges the investor interest on the money borrowed plus an additional service charge.

2.5.2 Margin Requirement

A margin purchase involves an element of risk for the broker. The shares they hold in street name form the collateral for the loan. If the value of the shares falls, then the collateral is reduced and the broker faces the risk that the borrower may default. To protect themselves against this, the broker insists that only a fraction of the investment be funded by borrowing. This is fraction termed the *initial margin requirement.*

The initial margin requirement, expressed as a percentage, is calculated by the formula

$$\text{Initial Margin Requirement} = \frac{\text{value financed by investor}}{\text{total value of investment}} \times 100. \tag{2.1}$$

This can be expressed alternatively by saying that the initial margin requirement is the minimum percentage of the investment that has to be financed by the investor. In the US, the Board of Governors of the Federal Reserve system has authorized that the initial margin must be at least 50%. Exchanges can impose a higher requirement than this, and this can be raised even further by brokers.

Example 11 *If the initial margin requirement is 60%, an investor must provide at least $6,000 of a $10,000 investment and the brokerage no more than $4,000.*

In the period following a margin purchase the value of the investment made will change. If the value falls far enough, then the collateral the brokerage is holding may no longer be sufficient to cover the loan. To guard against this, the brokerage calculates the value of the securities each day. This is called *marking to market*. From this is calculated the *actual margin* which is defined by

$$\text{Actual Margin} = \frac{\text{market value of assets - loan}}{\text{market values of assets}} \times 100. \tag{2.2}$$

The actual margin can rise or fall as the asset prices change.

Example 12 *Assume that a margin purchase of $10,000 has been made using $7,000 of the investor's own funds and $3,000 borrowed from the broker. If the value of the investment rises to $12,000 the actual margin is* $\frac{12{,}000-3{,}000}{12{,}000} \times 100 = 75\%$*. If instead the value of the investment falls to $6,000 the actual margin is* $\frac{6{,}000-3{,}000}{6{,}000} \times 100 = 50\%$*.*

A brokerage will require that the actual margin should not fall too far. If it did, there would be a risk that the investor may default and not pay off the loan. The *maintenance margin requirement* is the minimum value of the actual margin that is acceptable to the brokerage. The New York Stock Exchange imposes a maintenance margin of 25% and most brokers require 30% or more. If the actual margin falls below the maintenance margin, then a *margin call* is issued. A margin call requires that the investor must add further funds to the margin account or deposit additional assets. Either of these will raise the market value of assets in the account. Alternatively, part of the loan could be repaid. In any case, the action must be significant enough to raise the actual margin back above the maintenance margin.

Example 13 *Assume that a margin purchase of $12,000 has been made with a loan of $4,000. With a maintenance margin of 30%, the investor will receive a margin call when*

$$\frac{\textit{market value of assets - 4,000}}{\textit{market values of assets}} \times 100 < 30.$$

This is satisfied when the market value of assets is less than $5714.

2.5.3 Margin and Return

Buying on the margin has a both a benefit and a cost. Recall that the formula (1.1) defined the return as the increase in value of the investor as a percentage of the initial value. What changes when this formula is applied to a margin purchase is that the initial value of the investment is measured by the funds coming from the investor's own resources. With a margin purchase the quantity of the investor's funds is reduced for any given size of investment by the value of the funds borrowed from the brokerage. As the following example shows, this reduction magnifies the return obtained from the investment.

Example 14 *Consider an investment of \$5,000 made using a cash account. If the value of the investment rises to \$6,500 the cash return is*

$$\textit{Cash Return} = \frac{6,500 - 5,000}{5,000} \times 100 = 30\%. \tag{2.3}$$

Now consider the same investment using a margin account. Assume the initial margin is 60% so the investor provides \$3000 and borrows \$2000. With an interest rate of 10% charged on the loan the return is

$$\textit{Margin Return} = \frac{(6,500 - 2,000) - 3,000 - 0.1 \times 2,000}{3,000} \times 100 = 43\%. \tag{2.4}$$

Example 14 reveals the general property of a margin purchase which is that it raises the return above that of a cash purchase if the return is positive. This is because the return is calculated relative to the contribution of the investor which, due to the loan component, is less that for a cash purchase.

Margin purchases do have a downside though. As the following example shows, a margin purchase also magnifies negative returns.

Example 15 *Assume the value falls to \$4,000. The the return from a cash purchase is*

$$\textit{Cash Return} = \frac{4,000 - 5,000}{5,000} \times 100 = -20\%, \tag{2.5}$$

and the return on the margin purchase is

$$\textit{Margin Return} = \frac{(4,000 - 2,000) - 3,000 - 0.1 \times 2,000}{3,000} \times 100 = -40\%. \tag{2.6}$$

The conclusion from this analysis is that purchasing on the margin magnifies gains and losses. Because of this, it increases the risk of a portfolio. Informally, this suggests that a margin purchase should only really be considered when there is a strong belief that a positive return will be earned. Obviously, this conclusion can only be formally addressed using the techniques of portfolio analysis developed later.

2.6 Short Sales

A short sale is the sale of a security that an investor does not own. This can be achieved by borrowing shares from another investor. It is part of the role of a broker to organize such transactions and to ensure that the investor from whom the shares are borrowed does not suffer from any loss.

To provide the shares for a short sale, the broker either:

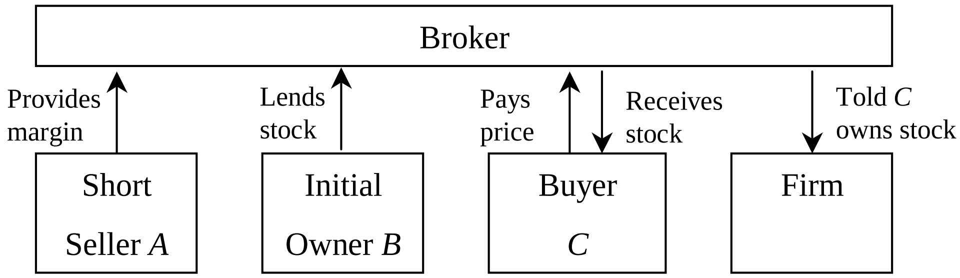


Figure 2.1: A Short Sale

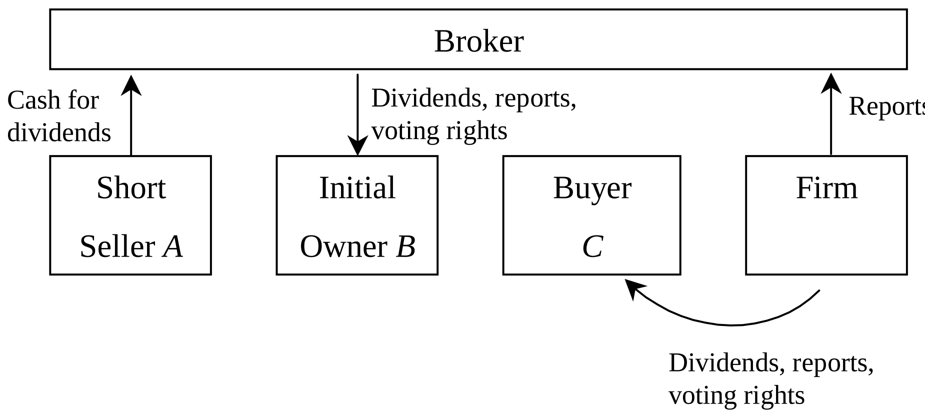


Figure 2.2: After the Short Sale

- Uses shares held in street name;

  or

- Borrows the shares from another broker.

Figures 2.1 and 2.2 illustrate a short sale. Investor A is the short-seller. The shares are borrowed from B and legally transferred to the buyer C. This is shown in Figure 2.1. To ensure that B does not lose from this short sale, A must pay any dividends that are due to B and the broker provides an annual report and voting rights. The report can come from the firm and the voting rights can be borrowed from elsewhere - either from other shares owned by the broker or from other brokers. Figure 2.2 illustrates this.

To close the transaction, the investor A must eventually purchase the shares and return them to B. A profit can only be made from the transaction if the shares can be purchased for less than they were sold. Short-selling is only used if prices are expected to fall.

There is a risk involved for the broker in organizing a short sale. If the investor defaults, the broker will have to replace the shares that have been borrowed. The short-seller must make an initial margin advance to the broker to cover them against this risk. This initial margin is calculated as a percentage of the value of the assets short-sold. The broker holds this in the investor's account until the short-sale is completed and the investor finally restores the shares to the initial owner.

Example 16 *Let 100 shares be short-sold at $20 per share. The total value of the transaction is $2000. If the initial margin requirement is 50%, the investor must deposit a margin of $1000 with the brokerage.*

To guard the brokerage against any losses through changes in the price of the stock, a maintenance margin is enforced. Thus a margin call is made if the actual margin falls below the maintenance margin. The actual margin is defined by

$$\text{Actual Margin} = \frac{\text{short sale proceeds + initial margin - value of stock}}{\text{value of stock}} \times 100, \tag{2.7}$$

where the value of stock is the market value of the stock that has been short-sold.

Example 17 *If the value of the shares in Example 16 rises to $2,500 the actual margin is*

$$\textit{Actual Margin} = \frac{2,000 + 1,000 - 2,500}{2,500} \times 100 = 20\%, \tag{2.8}$$

If instead they fall to $1,500 the actual margin becomes

$$\textit{Actual Margin} = \frac{2,000 + 1,000 - 1,500}{1,500} \times 100 = 100\%. \tag{2.9}$$

With a short sale actual margin rises as the value of the stock sold-short falls.

2.7 Summary

Trading is a necessary act in portfolio construction and management. Securities can be traded in a number of ways through brokers offering a range of service levels. These trading methods have been described, especially the process of short-selling which has important implications in the following chapters. The process of buying using a margin account has been shown to raise return, but also to increase potential losses. With the practical background of these introductory chapters it is now possible to begin the formalities of investment analysis.

Exercise 13 *A margin account is used to buy 200 shares on margin at $35 per share. $2000 is borrowed from the broker to complete the purchase. Determine the actual margin:*

a. When the purchase is made;
b. If the price of the stock rises to $45 per share;
c. If the price of the stock falls to $30 per share.

Exercise 14 *An investor buys 2000 shares at \$30 each. The initial margin requirement is 50% and the maintenance margin is 30%. Show that if the stock price falls to \$25, the investor will not receive a margin call. At what price will a margin call be received?*

Exercise 15 *600 shares are purchased on the margin at the beginning of the year for \$40 per share. The initial margin requirement was 55%. Interest of 10% was paid on the margin loan and no margin call was ever faced. A dividend of \$2 per share is received. Calculate the annual return if:*

a. The stock are sold for \$45 per share at the end of the year;

b. If the stock are sold for \$25 per share at the end of the year.

c. Calculate the return for (a) and (b) if the purchase had been made using cash instead of on the margin.

Exercise 16 *Using a margin account, 300 shares are short sold for \$30 per share. The initial margin requirement is 45%.*

a. If the price of the stock rises to \$45 per share, what is the actual margin in the account?

b. If the price of the stock falls to \$15 per share, what is the actual margin in the account?

Exercise 17 *Is it true that the potential loss on a short sale is infinite? What is the maximum return?*

Part II

Portfolio Theory

Chapter 3

Risk and Return

The first steps in investment analysis are to calculate the gains from an investment strategy and the risk involved in that strategy. Investment analysts choose to measure gains by using the concept of a return. This chapter will show how returns can be calculated in a variety of circumstances, both for individual assets and for portfolios. Looking back over the past performance of an investment the calculation of risk is just an exercise in computation. Given the data, the formulas will provide the answer. Where the process is interesting is when we look forward to what the return may be in the future. The challenge of investment analysis is that future returns can never be predicted exactly. The investor may have beliefs about what the return will be, but the market never fails to deliver surprises. Looking at future returns it is necessary to accommodate their unpredictability by determining the range of possible values for the return and the likelihood of each. This provides a value for the expected return from the investment. What remains is to determine just how uncertain the return is. The measure that is used to do this, the variance of return, is the analyst's measure of risk. Together the expected return and variance of alternative portfolios provide the information needed to compare investment strategies.

3.1 Introduction

At the heart of investment analysis is the observation that the market rewards those willing to bear risk. An investor purchasing an asset faces two potential sources of risk. The future price at which the asset can be sold may be unknown, as may the payments received from ownership of the asset. For a stock, both of these features are immediately apparent. The trading price of stocks changes almost continually on the exchanges. The payment from stocks comes in the form of a dividend. Although companies attempt to maintain some degree of constancy in dividends, they are only a discretionary payment rather than a commitment and their levels are subject to change.

These arguments may not seem to apply to bonds whose maturity value and payments seem certain. But bond prices do fluctuate so, although the maturity value is known, the value at any time before maturity is not. Furthermore, the maturity value is given in nominal terms whereas the real value is uncertain as inflation must be taken into account. The same argument also applies to the real value of the coupon payments. Finally, there is the risk of default or early redemption. Only the shortest term bonds issued by major governments can ever be regarded as having approximately certain payoffs.

In order to guide investment choice, an investor must be able to quantify both the reward for holding an asset and the risk inherent in that reward. They must also be aware of how the rewards and risks of individual assets interact when the assets are combined into a portfolio. This chapter shows how this is done.

3.2 Return

The measure of reward that is used in investment analysis is called the *return*. Although we focus on financial assets, the return can be calculated for any investment provided we know its initial value and its final value.

The return is defined as the increase in value over a given time period as a proportion of the initial value. The time over which the return is computed is often called the *holding period*. Returns can be written in the raw form just defined or, equally well, converted to percentages. All that matters in the choice between the two is that consistency is used throughout a set of calculations. If you start using percentages, they must used everywhere. The calculations here will typically give both.

The formula for calculating the return can now be introduced. Letting V_0 be the initial value of the investment and V_1 the final value at the end of the holding period, the return, r, is defined by

$$r = \frac{V_1 - V_0}{V_0}. \tag{3.1}$$

To express the return as a percentage the formula is modified to

$$r = \frac{V_1 - V_0}{V_0} \times 100. \tag{3.2}$$

Example 18 *An initial investment is made of \$10,000. One year later, the value of the investment has risen to \$12,500. The return on the investment is* $r = \frac{12500-10000}{10000} = 0.25$. *Expressed as a percentage,* $r = \frac{12500-10000}{10000} \times 100 = 25\%$.

It should be emphasized that the return is always measured relative to the holding period. The example used a year as the holding period, which is the conventional period over which most returns are expressed. For instance, interest rates on bonds and deposit accounts are usually quoted as an annual rate. The precise description of the return in the example is consequently that the return on the investment was 25% per year. Other time periods may be encountered such as a month, a week, or even a day. Detailed analysis of stock prices often employs daily returns.

Example 19 *An investment initially costs \$5,000. Three months later, the investment is sold for \$6,000. The return on the investment is* $r = \frac{6000-5000}{5000} \times 100 = 20\%$ *per three months.*

3.2.1 Stock Returns

The process for the calculation of a return can also be applied to stocks. When doing this it is necessary to take care with the payment of dividends since these must be included as part of the return. We first show how to calculate the return for a stock that does not pay a dividend and then extend the calculation to include dividends.

Consider a stock that pays no dividends for the holding period over which the return is to be calculated Assume that this period is one year. In the formula for the return, we take the initial value, V_0, to be the purchase price of the stock and the final value, V_1, to be its trading price one year later. If the initial price of the stock is $p(0)$ and the final price $p(1)$ then the return on the stock is

$$r = \frac{p(1) - p(0)}{p(0)}. \tag{3.3}$$

Example 20 *The price of Lastminute.com stock trading in London on May 29 2002 was £0.77. The price at close of trading on May 28 2003 was £1.39. No dividends were paid. The return for the year of this stock is given by*

$$r = \frac{1.39 - 0.77}{0.77} = 0.805 \text{ (80.5\%)}.$$

The method for calculating the return can now be extended to include the payment of dividends. To understand the calculation it needs to be recalled that the return is capturing the rate of increase of an investor's wealth. Since dividend payments are an addition to wealth, they need to be included in the calculation of the return. In fact, the total increase in wealth from holding the stock is the sum of its price increase plus the dividend received. So, in the formula for the return, the dividend is added to the final stock price.

Letting d denote the dividend paid by a stock over the holding period, this gives the formula for the return

$$r = \frac{p(1) + d - p(0)}{p(0)}. \tag{3.4}$$

Stocks in the US pay dividends four time per year and stock in the UK pay dividends twice per year. What there are multiple dividend payments during the holding period the value of d is the sum of these dividend payments.

Example 21 *The price of IBM stock trading in New York on May 29 2002 was $80.96. The price on May 28 2003 was $87.57. A total of $0.61 was paid in dividends over the year in four payments of $0.15, $0.15, $0.15 and $0.16. The return over the year on IBM stock was*

$$r = \frac{87.57 + 0.61 - 80.96}{80.96} = 0.089 \text{ (8.9\%).}$$

3.2.2 Portfolio Return

It was noted in the introduction that the definition of a return could be applied to any form of investment. So far it has only been applied to individual assets. We now show how the method of calculation can be applied to a portfolios of assets. The purchase of a portfolio is an example of an investment and consequently a return can be calculated.

The calculation of the return on a portfolio can be accomplished in two ways. Firstly, the initial and final values of the portfolio can be determined, dividends added to the final value, and the return computed. Alternatively, the prices and payments of the individual assets, and the holding of those assets, can be used directly.

Focussing first on the total value of the portfolio, if the initial value is V_0, the final value V_1, and dividends received are d, then the return is given by

$$r = \frac{V_1 + d - V_0}{V_0}. \tag{3.5}$$

Example 22 *A portfolio of 200 General Motors stock and 100 IBM stock is purchased for $20,696 on May 29 2002. The value of the portfolio on May 28 2003 was $15,697. A total of $461 in dividends was received. The return over the year on the portfolio is* $r = \frac{15697+461-20696}{20696} = -0.219$ *(-21.9%).*

The return on a portfolio can also be calculated by using the prices of the assets in the portfolio and the quantity of each asset that is held. Assume that an investor has constructed a portfolio composed of N different assets. The quantity held of asset i is a_i. If the initial price of asset i is $p_i(0)$ and the final price $p_i(1)$, then the initial value of the portfolio is

$$V_0 = \sum_{i=1}^{N} a_i p_i(0), \tag{3.6}$$

and the final value

$$V_1 = \sum_{i=1}^{N} a_i p_i(1). \tag{3.7}$$

If there are no dividends, then these can be used to calculate the return as

$$r = \frac{V_1 - V_0}{V_0} = \frac{\sum_{i=1}^{N} a_i p_i(1) - \sum_{i=1}^{N} a_i p_i(0)}{\sum_{i=1}^{N} a_i p_i(0)}. \tag{3.8}$$

Example 23 *Consider the portfolio of three stocks described in the table.*

Stock	Holding	Initial Price	Final Price
A	100	2	3
B	200	3	2
C	150	1	2

Example 24 *The return on the portfolio is*

$$\begin{aligned} r &= \frac{(100 \times 3 + 200 \times 2 + 150 \times 2) - (100 \times 2 + 200 \times 3 + 150 \times 1)}{100 \times 2 + 200 \times 3 + 150 \times 1} \\ &= 0.052 \textit{ (5.2\%).} \end{aligned}$$

This calculation can be easily extended to include dividends. If the dividend payment per share from stock i is denoted by d_i, the formula for the calculation of the return from a portfolio becomes

$$r = \frac{\sum_{i=1}^{N} a_i \left[p_i(1) + d_i\right] - \sum_{i=1}^{N} a_i p_i(0)}{\sum_{i=1}^{N} a_i p_i(0)} \tag{3.9}$$

Example 25 *Consider the portfolio of three stocks described in the table.*

Stock	Holding	Initial Price	Final Price	Dividend per Share
A	50	10	15	1
B	100	3	6	0
C	300	22	20	3

Example 26 *The return on the portfolio is*

$$\begin{aligned} r &= \frac{(50\,[15+1] + 100\,[6] + 300\,[20+3]) - (50\,[10] + 100\,[3] + 300\,[22])}{50\,[10] + 100\,[3] + 300\,[22]} \\ &= 0.122 \textit{ (12.2\%).} \end{aligned}$$

The calculation of the return can also be extended to incorporate short-selling of stock. Remember that short-selling refers to the act of selling an asset you do not own by borrowing the asset from another investor. In the notation used here, short-selling means you are indebted to the investor from whom the stock has been borrowed so that you effectively hold a negative quantity of the stock. For example, if you have gone short 200 shares of Ford stock, then the holding for Ford is given by -200. The return on a short sale can only be positive if the price of Ford stock falls. In addition, during the period of the short sale the short-seller is responsible for paying the dividend on the stock that they have borrowed. The dividends therefore count against the return since they are a payment made.

Example 27 *On June 3 2002 a portfolio is constructed of 200 Dell stocks and a short sale of 100 Ford stocks. The prices on these stocks on June 2 2003, and the dividends paid are given in the table.*

Stock	Initial Price ($)	Dividend ($)	Final Price ($)
Dell	26.18	0	30.83
Ford	17.31	0.40	11.07

Example 28 *The return over the year on this portfolio is*

$$\begin{aligned} r &= \frac{(200 \times 30.83 + [-100 \times 11.47]) - (200 \times 26.18 + [-100 \times 17.31])}{200 \times 26.18 + [-100 \times 17.31]} \\ &= 0.43 \ (43\%). \end{aligned}$$

3.2.3 Portfolio Proportions

The calculations of portfolio return so far have used the quantity held of each asset to determine the initial and final portfolio values. What proves more convenient in later calculations is to use the *proportion* of the portfolio invested in each asset rather then the total holding. The two give the same answer but using proportions helps emphasize that the returns (and the risks discussed later) depend on the mix of assets held, not on the size of the total portfolio.

The first step is to determine the proportion of the portfolio in each asset. If the value of the investment in asset i at the start of the holding period is V_0^i, then the proportion invested in asset i is defined by

$$X_i = \frac{V_0^i}{V_0}, \tag{3.10}$$

where V_0 is the initial value of the portfolio. By definition, these proportions must sum to 1. For a portfolio with N assets this can be seen from writing

$$\sum_{i=1}^{N} X_i = \frac{\sum_{i=1}^{N} V_0^i}{V_0} = \frac{V_0}{V_0} = 1. \tag{3.11}$$

Furthermore, if an asset i is short-sold then its proportion is negative, so $X_i < 0$. This again reflects the fact that short-selling is treated as a negative shareholding.

Example 29 *Consider the portfolio in Example 23. The initial value of the portfolio is* 950 *and the proportional holdings are*

$$X_A = \frac{200}{950},\ X_B = \frac{600}{950},\ X_C = \frac{150}{950}.$$

Example 30 *A portfolio consists of a purchase of 100 of stock A at \$5 each, 200 of stock B at \$3 each and a short-sale of 150 of stock C at \$2 each. The total value of the portfolio is*

$$V_0 = 100 \times 5 + 200 \times 3 - 150 \times 2 = 800.$$

The portfolio proportions are

$$X_A = \frac{5}{8},\ X_B = \frac{6}{8},\ X_C = -\frac{3}{8}.$$

Once the proportions have been calculated it is possible to evaluate the return on the portfolio. Using the proportions, the return is the weighted average of the returns on the individual assets. The return can be calculated using

$$r = \sum_{i=1}^{N} X_i r_i. \tag{3.12}$$

Example 31 *From the figures in Example 23, the returns on the stocks are*

$$r_A = \frac{3-2}{2} = \frac{1}{2},\ r_B = \frac{2-3}{3} = -\frac{1}{3},\ r_C = \frac{2-1}{1} = 1,$$

and from Example 29 the initial proportions in the portfolio are

$$X_A = \frac{200}{950},\ X_B = \frac{600}{950},\ X_C = \frac{150}{950}.$$

The return on the portfolio is therefore

$$r = \frac{200}{950} \times \left(\frac{1}{2}\right) + \frac{600}{950} \times \left(-\frac{1}{3}\right) + \frac{150}{950} \times (1) = 0.052 \textit{(52\%)}.$$

It is important to note that the portfolio proportions are calculated at the start of the holding period. If a series of returns is to be calculated over a number of holding periods, the proportions must be recomputed at the start of each of the holding periods. This is necessary to take into account variations in the relative values of the assets. Those that have relatively larger increases in value will gradually form a greater proportion of the portfolio.

Example 32 *A portfolio consists of two stocks, neither of which pays any dividends. The prices of the stock over a three year period and the holding of each is given in the table.*

Stock	Holding	$p(0)$	$p(1)$	$p(2)$	$p(3)$
A	100	10	15	12	16
B	200	8	9	11	12

Example 33 *The initial value of the portfolio is* $V_0 = 100 \times 10 + 200 \times 8 = 2600$*, so the portfolio proportions are*

$$X_A(0) = \frac{1000}{2600} = \frac{5}{13},\ X_B(0) = \frac{1600}{2600} = \frac{8}{13}.$$

The portfolio return over the first year is then

$$r = \frac{5}{13} \times \frac{15-10}{10} + \frac{8}{13} \times \frac{9-8}{8} = 0.269 \textit{ (26.9\%)}$$

At the start of the second year, the value of the portfolio is $V_1 = 100 \times 15 + 200 \times 9 = 3300$*. This gives the new portfolio proportions as*

$$X_A(1) = \frac{1500}{3300} = \frac{5}{11},\ X_B(0) = \frac{1800}{3300} = \frac{6}{11},$$

and return

$$r = \frac{5}{11} \times \left(\frac{12-15}{15} \right) + \frac{6}{11} \times \left(\frac{11-9}{9} \right) = 0.03 \text{ (3\%)}.$$

Finally, the proportions at the start of the third holding period are

$$X_A(2) = \frac{1200}{3400} = \frac{6}{17},\ X_B(2) = \frac{2200}{3400} = \frac{11}{17},$$

and the return is

$$r = \frac{6}{17} \times \frac{16-12}{12} + \frac{11}{17} \times \frac{12-11}{11} = 0.176 \text{ (17.6\%)}.$$

3.2.4 Mean Return

The examples have illustrated that over time the return on a stock or a portfolio may vary. The prices of the individual stocks will rise and fall, and this will cause the value of the portfolio to fluctuate. Once the return has been observed for a number of periods it becomes possible to determine the average, or *mean*, return. For the moment the mean return is taken just as an average of past returns. We discuss later how it can be interpreted as a predictor of what may be expected in the future.

If a return, on an asset or portfolio, is observed in periods $1, 2, 3, ..., T$, the mean return is defined as

$$\overline{r} = \sum_{t=1}^{T} \frac{r_t}{T}, \tag{3.13}$$

where r_t is the return in period t.

Example 34 *Consider the following returns observed over 10 years.*

Year	1	2	3	4	5	6	7	8	9	10
Return (%)	4	6	2	8	10	6	1	4	3	6

Example 35 *The mean return is*

$$\overline{r} = \frac{4+6+2+8+10+6+1+4+3+6}{10} = 5\%.$$

It should be emphasized that this is the mean return over a given period of time. For instance, the example computes the mean return per year over the previous ten years.

3.3 Variance and Covariance

The essential feature of investing is that the returns on the vast majority of financial assets are not guaranteed. The price of stocks can fall just as easily as they can rise, so a positive return in one holding period may become a negative in the next. For example, an investment in the shares of Yahoo! Inc. would have earned a return of 137% between October 2002 and September 2003. Three years later the return from October 2005 through to September 2006 was -31%. The following year the stock had a return of 2%. Changes of this magnitude in the returns in different holding periods are not exceptional.

It has already been stressed that as well as caring about the return on an asset or a portfolio and investor has to be equally concerned with the risk. What risk means in this context is the variability of the return across different holding periods. Two portfolios may have an identical mean return but can have very different amounts of risk. There are few (if any) investors who would knowingly choose to hold the riskier of the two portfolios.

A measure of risk must capture the variability. The standard measure of risk used in investment analysis is the *variance* of return (or, equivalently, its square root which is called the *standard deviation*). An asset with a return that never changes has no risk. For this asset the variance of return is 0. Any asset with a return that does vary will have a variance of return that is positive. The more risk is the return on an asset the larger is the variance of return.

When constructing a portfolio it is not just the risk on individual assets that matters but also the way in which this risk combines across assets to determine the portfolio variance. Two assets may be individually risky, but if these risks cancel when the assets are combined then a portfolio composed of the two assets may have very little risk. The risks on the two assets will cancel if a higher than average return on one of the assets always accompanies a lower than average return on the other. The measure of the way returns are related across assets is called the *covariance* of return. The covariance will be seen to be central to understanding portfolio construction.

The portfolio variance and covariance are now developed by first introducing the variance of return as a measure of the risk and then developing the concept of covariance between assets.

3.3.1 Sample Variance

The data in Table 3.1 detail the annual return on General Motors stock traded in New York over a 10 year period. Figure 3.1 provides a plot of this data. The variability of the return, from a maximum of 36% to a minimum of - 41%, can be clearly seen. The issue is how to provide a quantitative measure of this variability.

Year	93-94	94-95	95-96	96-97	97-98
Return %	36.0	-9.2	17.6	7.2	34.1
Year	98-99	99-00	00-01	01-02	02-03
Return %	-1.2	25.3	-16.6	12.7	-40.9

Table 3.1: Return on General Motors Stock 1993-2003

The sample variance is a single number that summarizes the extent of the variation in return. The process is to take the mean return as a measure of

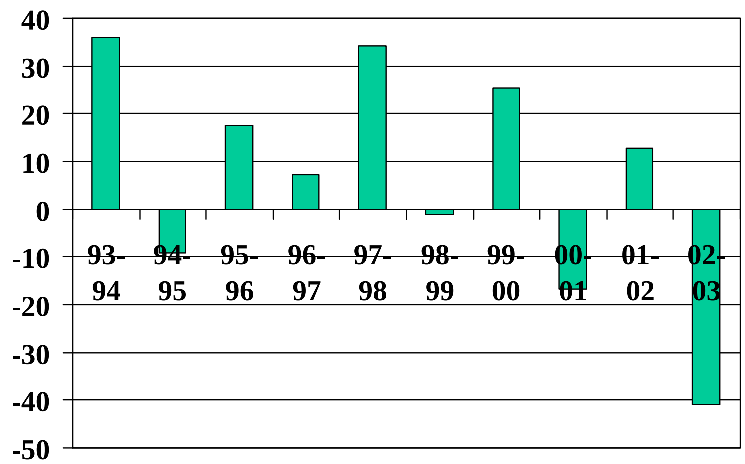


Figure 3.1: Graph of Return

the "normal" outcome. The difference between the mean and each observed return is then computed - this is termed the deviation from the mean. Some of these deviations from the mean are positive (in periods when the observed return is above the mean) and some are negative (when the observed return is below the mean). The deviations from the mean are then squared and these squares are summed. The average is then obtained by dividing by the number of observations.

With T observations, the sample variance just described is defined by the formula

$$\sigma^2 = \frac{1}{T}\sum_{t=1}^{T}(r_t - \overline{r})^2. \tag{3.14}$$

The sample standard deviation is the square root of the sample variance so

$$\sigma = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(r_t - \overline{r})^2}. \tag{3.15}$$

It should be noted that the sample variance and the sample standard deviation are always non-negative, so $\sigma^2 \geq 0$ and $\sigma \geq 0$. Only if every observation of the return is identical is the sample variance zero.

There is one additional statistical complication with the calculation of the variance. We can view the sample variance as being an estimate of the population variance of the return (meaning the true underlying value). The formula given in (3.14) for the sample variance produces an estimate of the population variance which is too low for small samples, that is when we have a small number of observations. (Although it does converge to the true value for large samples.)

Because of this, we say that it is a *biased estimator*. There is an alternative definition of the population variance which is unbiased. This is now described.

The unbiased estimator of the population variance is defined by

$$\sigma_{T-1}^2 = \frac{1}{T-1}\sum_{t=1}^{T}(r_t - \overline{r})^2, \tag{3.16}$$

with the unbiased estimator of the population standard deviation being

$$\sigma_{T-1} = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(r_t - \overline{r})^2}. \tag{3.17}$$

Comparing the formulas (3.14) and (3.16) it can be seen that the distinction between the two is simply whether the average value is found by dividing by T or $T-1$.

Either of these formulas is perfectly acceptable for a calculation of the sample variance. All that matters is that the same formula is used consistently. However, from this point onwards we will use division by T. It should be observed that as the number of observations increases, so T becomes large, the difference between dividing by T and by $T-1$ becomes ever less important. For very large values of T the two formulas provide approximately the same answer.

The next example calculates the sample variance of the return on General Motors stock using the data in Table 3.1.

Example 36 *For the returns on the General Motors stock, the mean return is*

$$\overline{r} = 6.5.$$

Using this value, the deviations from the mean and their squares are given by

Year	93-94	94-95	95-96	96-97	97-98
$r_t - \overline{r}$	29.5	-15.7	11.1	0.7	27.6
$(r_t - \overline{r})^2$	870.25	246.49	123.21	0.49	761.76
Year	98-99	99-00	00-01	01-02	02-03
$r_t - \overline{r}$	-7.7	18.8	-23.1	6.2	-47.4
$(r_t - \overline{r})^2$	59.29	353.44	533.61	38.44	2246.76

Example 37 *After summing and averaging, the variance is*

$$\sigma^2 = 523.4.$$

3.3.2 Sample Covariance

Every sports fan knows that a team can be much more (or less) than the sum of its parts. It is not just the ability of the individual players that matters but how they combine. The same is true for assets when they are combined into portfolios.

For the assets in a portfolio it is not just the variability of the return on each asset that matters but also the way returns vary across assets. A set of assets that are individually high performers need not combine well in a portfolio. Just like a sports team the performance of a portfolio is subtly related to the interaction of the component assets.

To see this point very clearly consider the example in Table 3.2. The table shows the returns on two stocks for the holding periods 2006 and 2007. Over the two years of data the mean return on each stock is 6 and the sample variances of the returns are $\sigma_A^2 = \sigma_B^2 = 16$. Both stocks have a positive sample variance so are individually risky investments.

Stock	Return in 2006	Return in 2007
A	10	2
B	2	10

Table 3.2

The outcome with respect to risk changes considerably when these stocks are combined into a portfolio. Consider a portfolio that has proportion $\frac{1}{2}$ of stock A and $\frac{1}{2}$ of stock B. With these proportions the return on the portfolio in 2006 was

$$r_p = \frac{1}{2}10 + \frac{1}{2}2 = 6, \tag{3.18}$$

and in 2007 the return was

$$r_p = \frac{1}{2}10 + \frac{1}{2}2 = 6. \tag{3.19}$$

This gives the sample mean return on the portfolio as

$$\overline{r}_p = \frac{6+6}{2} = 6. \tag{3.20}$$

This value is the same as for the individual stocks. The key point is the sample variance of the portfolio. Calculation of the sample variance gives

$$\sigma_p^2 = \frac{[6-6]^2 + [6-6]^2}{2} = 0, \tag{3.21}$$

so the portfolio has no risk. What the example shows is that assets that are individually risky can be combined into a portfolio in such a way that their variability cancels and the portfolio has a constant return.

The feature of the example that gives rise to this result is that across the two years a high return on one asset is accompanied by a low return on the other asset. Put another way, as we move between years an increase in return on one of the assets is met with an equal reduction in the return on the other. These changes exactly cancel when the assets are placed into a portfolio. This example teaches a fundamental lesson for portfolio theory: it is not just the variability of asset returns that matters but how the returns on the assets move relative to each other. In our example the moves are always in opposite directions and this was exploited in the design of the portfolio to eliminate variability in the return on the portfolio. The complete elimination of risk in the portfolio is an extreme feature of the example. The general property of portfolio construction is to obtain a reduction in risk by careful combination of assets.

In the same way that the variance is used to measure the variability of return of an asset or portfolio, we can also provide a measure of the extent to which the returns on different assets move relative to each other. To do this we need to define the *covariance* between the returns on two assets, which is the commonly-used measure of whether the returns move together or in opposite directions.

The covariance takes the deviations from the mean return for the two assets at time t, multiplies these together, sums over time and then averages. Hence, when both assets have returns above the mean, or both below the mean, a positive amount is contributed to the sum. Conversely, when one is below the mean and the other above, a negative amount is contributed to the sum. It is therefore possible for the covariance to be negative, zero or positive. A negative value implies the returns on the two assets tend to move in opposite directions (when one goes up, the other goes down) and a positive value that they tend to move in the same direction. A value of zero shows that, on average, there is no pattern of coordination in their returns.

To provide the formula for the covariance, let the return on asset A at time t be r_{At} and the mean return on asset A be $\overline{r}_A$. Similarly, the return on asset B at time t and the mean return are r_{Bt} and $\overline{r}_B$. The covariance of the return between these assets, denoted σ_{AB}, is

$$\sigma_{AB} = \frac{1}{T} \sum_{t=1}^{T} \left[r_{At} - \overline{r}_A \right] \left[r_{Bt} - \overline{r}_B \right]. \tag{3.22}$$

By definition, for any asset i it follows from comparison of formula (3.14) for the variance and (3.22) for the covariance that $\sigma_{ii} = \sigma_i^2$, so the covariance of the return between an asset and itself is its variance. Also, in the formula for the covariance it does not matter in which order we take asset A and asset B. This implies that the covariance of A with B is the same as the covariance of B with A or $\sigma_{AB} = \sigma_{BA}$.

Example 38 *The table provides the returns on three assets over a three-year period.*

Asset	Year 1	Year 2	Year 3
A	10	12	11
B	10	14	12
C	12	6	9

Example 39 *The mean returns are* $\overline{r}_A = 11, \overline{r}_B = 12, \overline{r}_C = 9$. *The covariance between A and B is*

$$\sigma_{AB} = \frac{1}{3} \left[[10 - 11] [10 - 12] + [12 - 11] [14 - 12] + [11 - 11] [12 - 12] \right] = 1.333,$$

while the covariance between A and C is

$$\sigma_{AC} = \frac{1}{3}\left[[10-11][12-9]+[12-11][6-9]+[11-11][9-9]\right] = -2,$$

and that between B and C

$$\sigma_{BC} = \frac{1}{3}\left[[10-12][12-9]+[14-12][6-9]+[12-12][9-9]\right] = -4.$$

For a set of assets the variances and covariances between the returns are often presented using a *variance-covariance matrix.* In a variance-covariance matrix the entries on the main diagonal are the variances while those off the diagonal are the covariances. Since $\sigma_{ij} = \sigma_{ji}$, only half the covariances need to be presented. Usually it is those below the main diagonal. For three assets A, B and C the variance-covariance matrix would be of the form

	A	B	C
A	σ_A^2		
B	σ_{AB}	σ_B^2	
C	σ_{AC}	σ_{BC}	σ_C^2

Example 40 *For the data in Example 38, the variance-covariance matrix is*

	A	B	C
A	0.666		
B	1.333	2.666	
C	−2	−4	6

3.4 Population Return and Variance

The concept of sample mean return that we have developed so far looks back over historical data to form an average of observed returns. The same is true of the formulation of the sample variance and sample covariance. The sample values are helpful to some degree to summarize the past behavior of returns but what is really needed for investment analysis are predictions about what may happen in the future. An investor needs this information to guide their current investment decisions. We now discuss the extent to which the sample returns and sample variances calculated on historical data can become predictions of future outcomes.

A conceptual framework for analyzing future returns can be constructed as follows: take an asset and determine the possible levels of return it may achieve, and the probability with which each level of return may occur. For instance, after studying its current business model we may feel that over the next year IBM stock can achieve a return of 2% with probability $\frac{1}{4}$, of 4% with probability $\frac{1}{2}$, and 6% with probability $\frac{1}{4}$. The possible payoffs, and the associated probabilities, capture both the essence of randomness in the return and the best view we can form on what might happen. It will be shown in this section how forming predictions in this way can be used to construct measures of risk and return.

Before proceeding to do this, it is worth reflecting on the link between this approach and the calculations of sample means and sample variances using historical data. At first sight, it would seem that the two are distinctly different processes. However, there is a clear link between the two. This link follows from adopting the perspective that the past data reflect the outcomes of earlier random events. The observed data then constitute random draws from the set of possible outcomes, with the rate of occurrence governed by a probability distribution.

Adopting the usual approach of statistical analysis, the historical data on observed returns are a sample from which we can obtain estimates of the true values. The mean return we have calculated from the sample of observed returns is a best estimate of the mean for the entire population of possible returns. The mean return for the population is often called the *expected return*. The name of "mean" is correctly used for the value calculated from the outcome of observation, while the name of "expected" is reserved for the statistical expectation. However, since the mean return is the best estimate of the expected return, the terms are commonly used interchangeably.

The same comments also apply to the sample variance and the sample covariance developed previously. They, too, are sample estimates of the population variance and covariance. This was the point behind the discussion of the population variance being a measure of the true variance. The issue of unbiasedness arose as a desirable property of the sample variance as an estimator of the population variance.

3.4.1 Expectations

The first step in developing this new perspective is to consider the formation of *expectations*. Although not essential for using the formulas developed below, it is important for understanding their conceptual basis.

Consider rolling a dice and observing the number that comes up. This is a simple random experiment that can yield any integer between 1 and 6 with probability $\frac{1}{6}$. The entire set of possible outcomes and their associated probabilities is then

$$\left\{1, \frac{1}{6}\right\}, \left\{2, \frac{1}{6}\right\}, \left\{3, \frac{1}{6}\right\}, \left\{4, \frac{1}{6}\right\}, \left\{5, \frac{1}{6}\right\}, \left\{6, \frac{1}{6}\right\}. \tag{3.23}$$

The expected value from this experiment can be thought of as the mean of the outcome observed if the experiment was repeated very many times. Let x denote the number obtained by observing a roll of the dice. This is one observation of the random variable X. The expected value of the random variable is denoted $E(X)$ and is given by the sum of possible outcomes, x, weighted by their probabilities. For the dice experiment the expected value is

$$E[X] = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6 = 3.5. \tag{3.24}$$

Notice the interesting feature that the expected value of 3.5 is not an outcome which will ever be observed - only the integers 1 to 6 ever appear. But this does not prevent 3.5 being the expected value.

Expressed in formal terms, assume we have an random event in which there are M possible outcomes. If outcome j is a value x_j, and occurs with probability π_j, then the expected value of the random variable X is

$$E[X] = \sum_{j=1}^{M} \pi_j x_j. \tag{3.25}$$

The idea of taking an expectation is not restricted to just the observed values of the random experiment. Return to the dice rolling example. For this experiment we may also be interested in the expected value of X^2. This can be computed as

$$E\left[X^2\right] = \frac{1}{6} \times 1 + \frac{1}{6} \times 4 + \frac{1}{6} \times 9 + \frac{1}{6} \times 16 + \frac{1}{6} \times 25 + \frac{1}{6} \times 36 = 15.167. \tag{3.26}$$

This expression is just the value of each possible outcome squared, multiplied by the probability and summed.

Observing this use of the expectation, we can recall that the variance is defined as the average value of the square of the deviation from the mean. This, too, is easily expressed as an expectation. For the dice experiment the expected value was 3.5 (which we can use as the value of the mean), so the expected value of the square of the deviation from the mean is

$$\begin{aligned} E\left[(X - E[X])^2\right] &= \frac{1}{6}[1 - 3.5]^2 + \frac{1}{6}[2 - 3.5]^2 + \frac{1}{6}[3 - 3.5]^2 \\ &+ \frac{1}{6}[4 - 3.5]^2 + \frac{1}{6}[5 - 3.5]^2 + \frac{1}{6}[6 - 3.5]^2 = 2.9167. \end{aligned} \tag{3.27}$$

This is the population variance of the observed value of the dice rolling experiment.

3.4.2 Expected Return

The expectation can now be employed to evaluate the expected return on an asset and a portfolio. This is achieved by introducing the idea of *states of the world*. A state of the world summarizes all the information that is relevant for the future return of an asset, so the set of states describes all the possible different future financial environments that may arise. Of course, only one of these states will actually be realized but when looking forward we do not know which one. These states of the world are the analysts way of thinking about, and modelling, what generates the randomness in asset returns.

Let there be M states of the world. If the return on an asset in state j is r_j, and the probability of state j occurring is π, then the *expected return* on asset i is

$$E[r] = \pi_1 r_1 + ... + \pi_M r_M, \tag{3.28}$$

or, using the same notation as for the mean,

$$\bar{r} = \sum_{j=1}^{M} \pi_j r_j. \tag{3.29}$$

Example 41 *The temperature next year may be hot, warm or cold. The returns to stock in a food production company in each of these states are given in the table.*

State	Hot	Warm	Cold
Return	10	12	18

Example 42 *If each states is expected to occur with probability $\frac{1}{3}$, the expected return on the stock is*

$$E[r] = \frac{1}{3}10 + \frac{1}{3}12 + \frac{1}{3}18 = 13.333.$$

This method of calculating the expected return can be generalized to determine the expected return on a portfolio. This is done by observing that the expected return on a portfolio is the weighted sum of the expected returns on each of the assets in the portfolio.

To see this, assume we have N assets and M states of the world. The return on asset i in state j is r_{ij} and the probability of state j occurring is $\pi_{j.}$ Let X_i be the proportion of the portfolio invested in asset i. The return on the portfolio in state j is found by weighting the return on each asset by its proportion in the portfolio then summing

$$r_{Pj} = \sum_{i=1}^{N} X_i r_{ij.} \tag{3.30}$$

The expected return on the portfolio is found from the returns in the separate states and the probabilities so

$$E[r_P] = \pi_1 r_{P1} + ... + \pi_M r_{PM}. \tag{3.31}$$

The return on the portfolio in each state can now be replaced by its definition in terms of the returns on the individual assets to give

$$E[r_P] = \sum_{i=1}^{N} \pi_1 X_i r_{i1} + ... + \sum_{i=1}^{N} \pi_M X_i r_{iM}, \tag{3.32}$$

Collecting the terms for each asset

$$E[r_P] = \sum_{i=1}^{N} X_i [\pi_1 r_{i1} + ... + \pi_M r_{iM}], \tag{3.33}$$

which can be written in brief as

$$\bar{r}_P = \sum_{i=1}^{N} X_i \bar{r}_i. \tag{3.34}$$

As we wanted to show, the expected return on the portfolio is the sum of the expected returns on the assets multiplied by the proportion of each asset in the portfolio.

Example 43 *Consider a portfolio composed of two assets A and B. Asset A constitutes 20% of the portfolio and asset B 80%. The returns on the assets in the 5 possible states of the world and the probabilities of those states are given in the table.*

State	1	2	3	4	5
Probability	0.1	0.2	0.4	0.1	0.2
Return on A	2	6	1	9	2
Return on B	5	1	0	4	3

Example 44 *The expected return on asset A is*

$$\bar{r}_A = 0.1 \times 2 + 0.2 \times 6 + 0.4 \times 1 + 0.1 \times 9 + 0.2 \times 2 = 3.1,$$

and that on asset B is

$$\bar{r}_B = 0.1 \times 5 + 0.2 \times 1 + 0.4 \times 0 + 0.1 \times 4 + 0.2 \times 3 = 1.7.$$

The expected portfolio return is

$$\bar{r}_P = 0.2 \times 3.1 + 0.8 \times 1.7 = 1.98.$$

Notice that the same result is obtained by writing

$$\bar{r}_P = 0.1 \times (0.2 \times 2 + 0.8 \times 5) + 0.2 \times (0.2 \times 6 + 0.8 \times 1) + 0.4 \times (0.2 \times 1 + 0.8 \times 0) + 0.1 \times (0.2 \times 9 + 0.8 \times 4) + 0.2 \times (0.2 \times 2 + 0.8 \times 3) = 1.98.$$

3.4.3 Population Variance

The population variance mirrors the interpretation of the sample variance as being the average of the square of the deviation from the mean. But where the sample variance found the average by dividing by the number of observations (or one less than the number of observations), the population variance averages by weighting each squared deviation from the mean by the probability of its occurrence.

In making this calculation we follow the procedure introduced for the population mean of:

(i) Identifying the different states of the world;
(ii) Determining the return in each state;
(ii) Setting the probability of each state being realized.

We begin with the definition of the population variance of return for a single asset. The population variance is expressed in terms of expectations by

$$\sigma^2 = E\left[(r - E[r])^2\right]. \tag{3.35}$$

In this formula $E[r]$ is the population mean return. This is the most general expression for the variance which we refine into a form for calculation by making explicit how the expectation is calculated.

To permit calculation using this formula the number of states of the world, their returns and the probability distribution of states, must be specified. Let there be M states, and denote the return on the asset in state j by r_j. If the probability of state j occurring is π_j, the population variance of the return on the asset can be written as

$$\sigma^2 = \sum_{j=1}^{M} \pi_j \left[[r_j - \bar{r}]^2\right]. \tag{3.36}$$

Since it is a positively weighted sum of squares the population variance is always non-negative. It can be zero, but only if the return on the asset is the same in every state.

The population standard deviation is given by the square root of the variance, so

$$\sigma = \sqrt{\sum_{j=1}^{M} \pi_j \left[[r_j - \bar{r}]^2\right]}. \tag{3.37}$$

Example 45 *The table provides data on the returns on a stock in the five possible states of the world and the probabilities of those states.*

State	1	2	3	4	5
Return	5	2	-1	6	3
Probability	.1	.2	.4	.1	.2

Example 46 *For this data, the population variance is*

$$\begin{aligned}\sigma^2 &= .1[5-3]^2 + .2[2-3]^2 + .4[-1-3]^2 + .1[6-3]^2 + .2[3-3]^2 \\ &= 7.9.\end{aligned}$$

3.4.4 Population Covariance

The sample covariance was introduced as a measure of the relative movement of the returns on two assets. It was positive if the returns on the assets tended to move in the same direction, and negative if they had a tendency to move in opposite directions. The population covariance extends this concept to the underlying model of randomness in asset returns.

For two assets A and B, the population covariance, σ_{AB}, is defined by

$$\sigma_{AB} = E\left[(r_A - E[r_A])(r_B - E[r_B])\right]. \tag{3.38}$$

The expression of the covariance using the expectation provides the most general definition. This form is useful for theoretical derivations but needs to be given a more concrete form for calculations.

Assume there are M possible states of the world with state j having probability π_j. Denote the return to asset A in state j by r_{Aj} and the return to asset B in state j by r_{Bj}. The population covariance between the returns on two assets A and B can be written as

$$\sigma_{AB} = \sum_{j=1}^{M} \pi_j \left[r_{Aj} - \bar{r}_A \right] \left[r_{Bj} - \bar{r}_B \right], \tag{3.39}$$

where $\bar{r}_A$ and $\bar{r}_B$ are the expected returns on the two assets.

The population covariance may be positive or negative. A negative covariance arises when the returns on the two assets tend to move in opposite directions, so that if asset A has a return above its mean ($r_{Aj} - \bar{r}_A > 0$) then asset B has a return below its mean ($r_{Bj} - \bar{r}_B < 0$) and *vice versa*. A positive covariance arises if the returns on the assets tend to move in the same direction, so both are either above the mean or both are below the mean.

Example 47 *Consider the returns on three stocks in the following table. Assume the probability of the states occurring are:* $\pi_1 = \frac{1}{2}$, $\pi_2 = \frac{1}{4}$, $\pi_3 = \frac{1}{4}$.

State	1	2	3
Stock A	7	2	6
Stock B	8	1	6
Stock C	3	7	2

Example 48 *The mean returns on the stocks can be calculated as* $\bar{r}_A = 5$, $\bar{r}_B = 5$ *and* $\bar{r}_C = 4$. *The variance of return for the three stocks can be found as*

$$\begin{aligned}
\sigma_A^2 &= \frac{1}{2}(7-5)^2 + \frac{1}{4}(2-5)^2 + \frac{1}{4}(6-5)^2 = 4.5, \\
\sigma_B^2 &= \frac{1}{2}(8-5)^2 + \frac{1}{4}(1-5)^2 + \frac{1}{4}(6-5)^2 = 8.75, \\
\sigma_C^2 &= \frac{1}{2}(3-4)^2 + \frac{1}{4}(7-4)^2 + \frac{1}{4}(2-4)^2 = 3.75.
\end{aligned}$$

The covariances between the returns are

$$\begin{aligned}
\sigma_{AB} &= \frac{1}{2}(7-5)(8-5) + \frac{1}{4}(2-5)(1-5) + \frac{1}{4}(6-5)(6-5) = 6.25, \\
\sigma_{AC} &= \frac{1}{2}(7-5)(3-4) + \frac{1}{4}(2-5)(7-4) + \frac{1}{4}(6-5)(2-4) = -3.75, \\
\sigma_{BC} &= \frac{1}{2}(8-5)(3-4) + \frac{1}{4}(1-5)(7-4) + \frac{1}{4}(6-5)(2-4) = -5.0.
\end{aligned}$$

These can be summarized in the variance-covariance matrix

$$\begin{bmatrix} 4.5 & & \\ 6.25 & 8.75 & \\ -3.75 & -5.0 & 3.75 \end{bmatrix}.$$

3.5 Portfolio Variance

The calculations of the variance of the return on an asset and of the covariance of returns between two assets are essential ingredients to the determination of the variance of a portfolio. It has already been shown how a portfolio may have a very different variance from that of the assets from which it is composed. Why this occurs is one of the central lessons of investment analysis. The fact that it does has very significant implications for investment analysis

The variance of the return on a portfolio can be expressed in the same way as the variance on an individual asset. If the return on the portfolio is denoted by r_P and the mean return by $\bar{r}_P$, the portfolio variance, σ_P^2, is

$$\sigma_P^2 = E\left[(r_P - \bar{r}_P)^2\right]. \tag{3.40}$$

The aim now is to present a version of this formula from which the variance can be calculated. Achieving this aim should also lead to an understanding of how the variance of the return on the portfolio is related to the variances of the returns on the individual assets and the covariances between the returns on the assets.

The analysis begins by studying the variance of a portfolio with just two assets. The result obtained is then extended to portfolios with any number of assets.

3.5.1 Two Assets

Consider a portfolio composed of two assets, A and B, in proportions X_A and X_B. Using the definition of the population variance, the variance of the return on the portfolio is given by the expected value of the deviation of the return from the mean return squared.

The analysis of portfolio return has shown that $r_P = X_A r_A + X_B r_B$ and $\bar{r}_P = X_A \bar{r}_A + X_B \bar{r}_B$. These expressions can be substituted into the definition of the variance of the return on the portfolio to write

$$\sigma_P^2 = E\left[([X_A r_A + X_B r_B] - [X_A \bar{r}_A + X_B \bar{r}_B])^2\right]. \tag{3.41}$$

Collecting together the terms relating to asset A and the terms relating to asset B gives

$$\sigma_P^2 = E\left[(X_A [r_A - \bar{r}_A] + X_B [r_B - \bar{r}_B])^2\right]. \tag{3.42}$$

Squaring the term inside the expectation

$$\sigma_P^2 = E\left[X_A^2 [r_A - \bar{r}_A]^2 + X_B^2 [r_B - \bar{r}_B]^2 + 2X_A X_B [r_A - \bar{r}_A][r_B - \bar{r}_B]\right]. \tag{3.43}$$

The expectation of a sum of terms is equal to the sum of the expectations of the individual terms. This allows that variance to be broken down into separate expectations

$$\sigma_P^2 = E\left[X_A^2\left[r_A-\bar{r}_A\right]^2\right]+E\left[X_B^2\left[r_B-\bar{r}_B\right]^2\right] + E\left[2X_AX_B\left[r_A-\bar{r}_A\right]\left[r_B-\bar{r}_B\right]\right]. \tag{3.44}$$

The portfolio proportions can then be extracted from the expectations because they are constants. This gives

$$\sigma_P^2 = X_A^2E\left[\left[r_A-\bar{r}_A\right]^2\right]+X_B^2E\left[r_B-\bar{r}_B\right]^2 + 2X_AX_BE\left[\left[r_A-\bar{r}_A\right]\left[r_B-\bar{r}_B\right]\right]. \tag{3.45}$$

The first expectation in this expression is the variance of return on asset A, the second expectation is the variance of return on asset B, and the third expectation is the covariance of the returns of A and B. Employing these observation allows the variance of the return on a portfolio of two assets, A and B, to be written succinctly as

$$\sigma_P^2 = X_A^2\sigma_A^2 + X_B^2\sigma_B^2 + 2X_AX_B\sigma_{AB}. \tag{3.46}$$

The expression in (3.46) can be used to calculate the variance of the return on the portfolio given the shares of the two assets in the portfolio, the variance of returns of the two assets, and the covariance. The result has been derived for the population variance (so the values entering would be population values) but can be used equally well to calculate the sample variance of the return on the portfolio using sample variances and sample covariance.

Example 49 *Consider two assets A and B described by the variance-covariance matrix*

$$\begin{bmatrix} 4 & \\ 2 & 8 \end{bmatrix}.$$

The variance of a portfolio consisting of $\frac{1}{4}$ asset A and $\frac{3}{4}$ asset B is given by

$$\sigma_P^2 = \left(\frac{1}{4}\right)^2 4 + \left(\frac{3}{4}\right)^2 8 + 2\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)2 = 5.5.$$

Example 50 *Consider two assets C and D described by the variance-covariance matrix*

$$\begin{bmatrix} 6 & \\ -3 & 9 \end{bmatrix}.$$

The variance of a portfolio consisting of $\frac{2}{3}$ asset C and $\frac{1}{3}$ asset D is given by

$$\sigma_P^2 = \left(\frac{2}{3}\right)^2 6 + \left(\frac{1}{3}\right)^2 9 + 2\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)(-3) = 2\frac{1}{3}.$$

It can be seen from formula (3.46) for the variance of return on a portfolio that if the covariance between the two assets is negative, the portfolio variance is reduced. This observation is emphasized in the example by the variance of the portfolio of assets C and D being much lower that the portfolio of assets A and B. The variance-reducing effect of combining assets whose returns have a negative covariance is a fundamental result for investment analysis. It provides a clear insight into how the process for constructing portfolios can reduce the risk involved in investment.

3.5.2 Correlation Coefficient

The variance of the return on a portfolio can be expressed in an alternative way that is helpful in the analysis of the next chapter. The covariance has already been described as an indicator of the tendency of the returns on two assets to move in the same direction (either up or down) or in opposite directions. Although the sign of the covariance (whether it is positive or negative) indicates this tendency, the value of the covariance does not in itself reveal how strong the relationship is. For instance, a given value of covariance could be generated by two assets that each experience large deviations from the mean but only have a weak relationship between their movements or by two assets whose returns are very closely related but individually do not vary much from their means.

In order to determine the strength of the relationship it is necessary to measure the covariance relative to the deviation from the mean experienced by the individual assets. This is achieved by using the *correlation coefficient* which relates the standard deviations and covariance. The correlation coefficient between the return on asset A and the return on asset B is defined by

$$\rho_{AB} = \frac{\sigma_{AB}}{\sigma_A \sigma_B}. \tag{3.47}$$

The value of the correlation coefficient satisfies $-1 \leq \rho_{AB} \leq 1$.

A value of $\rho_{AB} = 1$ indicates *perfect positive correlation*: the returns on the two assets always move in unison. Interpreted in terms of returns in different states of the world, perfect positive correlation says that if the return on one asset is higher in state j than it is in state k, then so is the return on the other asset. Conversely, $\rho_{AB} = -1$ indicates *perfect negative correlation*: the returns on the two assets always move in opposing directions, so if the return on one asset is higher in state j than it is in state k, then the return on the other asset is lower in state j than in state k.

Using the correlation coefficient, the variance of the return of a portfolio can be written as

$$\sigma_P^2 = X_A^2 \sigma_A^2 + X_B^2 \sigma_B^2 + 2 X_A X_B \rho_{AB} \sigma_A \sigma_B. \tag{3.48}$$

It can be seen from this formula that a negative correlation coefficient reduces the overall variance of the portfolio.

Example 51 *A portfolio is composed of $\frac{1}{2}$ of asset A and $\frac{1}{2}$ of asset B. Asset A has a variance of 25 and asset B a variance of 16. The covariance between* *the returns on the two assets is 10. The correlation coefficient is*

$$\rho_{AB} = \frac{10}{5 \times 4} = 0.5,$$

and the variance of return on the portfolio is

$$\sigma_P^2 = \left(\frac{1}{2}\right)^2 25 + \left(\frac{1}{2}\right)^2 16 + 2\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) 0.5 \times 25 \times 16 = 110.25.$$

3.5.3 General Formula

The formula to calculate the variance of the return on a portfolio can now be extended to accommodate any number of assets. This extension is accomplished by noting that the formula for the variance of the return on a portfolio involves the variance of each asset plus its covariance with every other asset.

For N assets in proportions X_i, $i = 1, ..., N$, the variance is therefore given by

$$\sigma_P^2 = \sum_{i=1}^{N} \left[X_i^2 \sigma_i^2 + \sum_{\substack{k=1 \\ k \neq i}}^{N} X_i X_k \sigma_{ik} \right]. \tag{3.49}$$

It should be confirmed that when $N = 2$ this reduces to (3.46). The presentation of the formula can be simplified by using the fact that σ_{ii} is identical to σ_i^2 to write

$$\sigma_P^2 = \sum_{i=1}^{N} \sum_{k=1}^{N} X_i X_k \sigma_{ik}. \tag{3.50}$$

This formula can also be expressed in terms of the correlation coefficients.

The significance of this formula is that it provides a measure of the risk of any portfolio, no matter how many assets are included. Conceptually, it can be applied even to very large (meaning thousands of assets) portfolios. All the information that is necessary to do this are the proportionate holdings of the assets and the variance-covariance matrix. Later chapters consider how this informational requirement can be reduced even further.

Example 52 *A portfolio consists of three assets, A, B, and C. The portfolio proportions are $X_A = \frac{1}{6}$, $X_B = \frac{1}{2}$, and $X_C = \frac{1}{3}$. The variance-covariance matrix is*

$$\begin{bmatrix} 3 & & \\ 4 & 12 & \\ 2 & -1 & 9 \end{bmatrix}.$$

The formula for the variance of the portfolio is

$$\sigma_P^2 = X_A^2 \sigma_A^2 + X_B^2 \sigma_B^2 + X_C^2 \sigma_C^2 + 2X_A X_B \sigma_{AB} + 2X_A X_C \sigma_{AC} + 2X_B X_C \sigma_{BC}.$$

Using the data describing the portfolio

$$\begin{aligned}\sigma_P^2 &= \left(\frac{1}{6}\right)^2 \sigma_A^2 + \left(\frac{1}{2}\right)^2 \sigma_B^2 + \left(\frac{1}{3}\right)^2 \sigma_C^2 + 2\frac{1}{6}\frac{1}{2}\sigma_{AB} + 2\frac{1}{6}\frac{1}{3}\sigma_{AC} + 2\frac{1}{2}\frac{1}{3}\sigma_{BC} \\ &= \frac{1}{36}\sigma_A^2 + \frac{1}{4}\sigma_B^2 + \frac{1}{9}\sigma_C^2 + \frac{1}{6}\sigma_{AB} + \frac{1}{9}\sigma_{AC} + \frac{1}{3}\sigma_{BC}.\end{aligned}$$

Substituting in from the variance-covariance matrix

$$\begin{aligned}\sigma_P^2 &= \left(\frac{1}{36}\right)3 + \left(\frac{1}{4}\right)12 + \left(\frac{1}{9}\right)9 + \left(\frac{1}{6}\right)4 + \left(\frac{1}{9}\right)2 + \left(\frac{1}{3}\right)(-1) \\ &= 4.6389.\end{aligned}$$

Example 53 *A portfolio consists of three assets, A, B, and C. The portfolio proportions are* $X_A = \frac{1}{4}$, $X_B = \frac{1}{4}$, *and* $X_C = \frac{1}{2}$. *The variances of the returns on the individual assets are* $\sigma_A^2 = 16$, $\sigma_B^2 = 25$, *and* $\sigma_C^2 = 36$. *The correlation coefficients between the returns are* $\rho_{AB} = 0.5$, $\rho_{BC} = 0.25$, *and* $\rho_{AC} = -0.75$. *The formula for the variance of the portfolio is*

$$\sigma_P^2 = X_A^2\sigma_A^2 + X_B^2\sigma_B^2 + X_C^2\sigma_C^2 + 2X_AX_B\sigma_A\sigma_B\rho_{AB} + 2X_AX_C\sigma_A\sigma_C\rho_{AC} + 2X_BX_C\sigma_B\sigma_C\rho_{BC}.$$

For the data describing the assets and the portfolio

$$\begin{aligned}\sigma_P^2 &= \left(\frac{1}{4}\right)^2 16 + \left(\frac{1}{4}\right)^2 25 + \left(\frac{1}{2}\right)^2 36 + 2\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)(4)(5)(0.5) \\ &\quad +2\left(\frac{1}{4}\right)\left(\frac{1}{2}\right)(4)(6)(-0.75) + 2\left(\frac{1}{4}\right)\left(\frac{1}{2}\right)(5)(6)(0.25) \\ &= 10.188.\end{aligned}$$

3.5.4 Effect of Diversification

As an application of the formula for the variance of the return of a portfolio this section considers the effect of diversification. Diversification means purchasing a larger number of different assets. It is natural to view diversification as a means of reducing risk because in a large portfolio the random fluctuations of individual assets will have a tendency to cancel out.

To formalize the effect of diversification, consider holding N assets in equal proportions. This implies that the portfolio proportions satisfy $X_i = \frac{1}{N}$ for all assets $i = 1, ..., N$. From (3.49), the variance of this portfolio is

$$\sigma_P^2 = \sum_{i=1}^{N}\left[\left[\frac{1}{N}\right]^2 \sigma_i^2 + \sum_{k=1,k\neq i}^{N}\left[\frac{1}{N}\right]^2 \sigma_{ik}\right]. \tag{3.51}$$

Observe that there are N terms in the first summation and $N[N-1]$ in the second. This suggests extracting a term from each summation to write the variance as

$$\sigma_P^2 = \left[\frac{1}{N}\right]\sum_{i=1}^{N}\left[\frac{1}{N}\right]\sigma_i^2 + \left[\frac{N-1}{N}\right]\sum_{i=1}^{N}\sum_{k=1,k\neq i}^{N}\left[\frac{1}{[N-1]N}\right]\sigma_{ik}. \tag{3.52}$$

Now define the mean of the variances of the N assets in the portfolio by

$$\bar{\sigma}_a^2 = \sum_{i=1}^{N} \left[\frac{1}{N}\right] \sigma_i^2, \tag{3.53}$$

and the mean covariance between all pairs of assets in the portfolio by

$$\bar{\sigma}_{ab} = \sum_{i=1}^{N} \sum_{k=1,k\neq i}^{N} \left[\frac{1}{[N-1]\,N}\right] \sigma_{ik}. \tag{3.54}$$

Using these definitions, the variance of the return on the portfolio becomes

$$\sigma_P^2 = \left[\frac{1}{N}\right] \bar{\sigma}_a^2 + \left[\frac{N-1}{N}\right] \bar{\sigma}_{ab}. \tag{3.55}$$

This formula applies whatever the number of assets (but the mean variance and mean covariance change in value as N changes).

Diversification means purchasing a broader range of assets which in the present context is reflected in an increase in N. The extreme of diversification occurs as the number of assets in the portfolio is increased without limit. Formally, this can be modelled by letting $N \to \infty$ and determining the effect on the variance of the return on the portfolio.

It can be seen from (3.55) that as $N \to \infty$ the first term will converge to zero (we are dividing the mean value by an ever increasing value of N) and the second term will converge to $\bar{\sigma}_{ab}$ (because as N increases $\frac{N-1}{N}$ tends to 1). Therefore, at the limit of diversification

$$\sigma_P^2 \to \bar{\sigma}_{ab}. \tag{3.56}$$

This result shows that in a well-diversified portfolio only the covariance between assets counts for portfolio variance. In other words, the variance of the individual assets can be eliminated by diversification - which confirms the initial perspective on the consequence of diversification.

3.6 Summary

The most basic information about assets is captured in their mean and variance which are used by analysts as the measures of return and risk. This chapter has shown how the sample return, sample variance and sample covariance can be calculated from data on individual assets. It has also shown how these can be combined into measures of risk and return for portfolios, including portfolios with short-selling of one or more assets.

These ideas were then extended to the calculation of population mean, variance and covariance. The calculation of population values was based upon the idea that the sample data was a random draw from an underlying population. Following this approach lead to the concept an expected value. The concepts involved in calculating population values capture the very essence of unpredictability in financial data.

Finally, the chapter applied the concept of the population variance as an expectation to calculate the variance of return on a portfolio. The importance of the covariance between the returns on the assets for this variance was stressed. This was emphasized further by presenting the variance in terms of the correlation coefficient and by demonstrating how diversification reduced the portfolio variance to the average of the covariances between assets in the portfolio.

Exercise 18 *A 1969 Jaguar E-type is purchased at the beginning of January 2002 for $25000. At the end of December 2002 it is sold for $30000.*

a. Given these figures, what was the return to the investment in the Jaguar?

b. Now assume the car was entered in a show and won a $500 prize. What does the return now become?

c. If in addition, it cost $300 to insure the car and $200 to service it, what is the return?

Exercise 19 *The following prices are observed for the stock of Fox Entertainment Group Inc.*

Date	June 00	June 01	June 02	June 03
Price	26.38	28.05	25.15	28.60

Exercise 20 *No dividend was paid. Calculate the mean return and variance of Fox stock.*

Exercise 21 *The returns on a stock over the previous ten years are as given in the table.*

Year	1	2	3	4	5	6	7	8	9	10
Return (%)	1	-6	4	12	2	-1	3	8	2	12

Exercise 22 *Determine the mean return on the stock over this period and its variance.*

Exercise 23 *The prices of three stocks are reported in the table.*

	June 00	June 01	June 02	June 03
Brunswick Corporation	16.56	24.03	28.00	23.00
Harley-Davidson Inc.	8.503	47.08	51.27	43.96
Polaris Industries Partners	31.98	45.80	65.00	63.04

Exercise 24 *During these years, the following dividends were paid*

	00-01	01-02	02-03
Brunswick Corporation	0.52	0.26	0.50
Harley-Davidson Inc.	0.12	0.09	0.12
Polaris Industries Partners	0.94	0.53	1.18

Exercise 25 *a. For each stock, calculate the return for each year and the mean return.*

b. Compute the return to a portfolio consisting of 100 Brunswick Corporation stock and 200 Harley-Davidson Inc. stock for each year.

c. For a portfolio of 100 of each of the stock, calculate the portfolio proportions at the start of each holding period. Hence compute the return to the portfolio.

Exercise 26 *For the data in Exercise 23 calculate the variance of return for each stock and the covariances between the stock. Discuss the resulting covariances paying particular attention to the market served by the companies. (If you do not know these companies, descriptions of their activities can be found on finance.yahoo.com.)*

Exercise 27 *Assume that there are 2 stocks and 5 states of the world. Each state can occur with equal probability. Given the returns in the following table, calculate the expected return and variance of each stock and the covariance between the returns. Hence find the expected return and variance of a portfolio with equal proportions of both stock. Explain the contrast between the variance of each stock and the portfolio variance.*

	State 1	State 2	State 3	State 4	State 5
Stock A	5	7	1	8	3
Stock B	9	6	5	4	8

Exercise 28 *Given the following variance-covariance matrix for three securities, calculate the standard deviation of a portfolio with proportional investments in the assets $X_A = 0.2$, $X_B = 0.5$ and $X_C = 0.3$.*

	Security A	Security B	Security C
Security A	24		
Security B	12	32	
Security C	10	-8	48

Exercise 29 *Consider the following standard deviations and correlation coefficients for three stocks.*

		Correlation	with	stock
Stock	σ	A	B	C
A	9	1	0.75	-0.5
B	6	0.75	1	0.2
C	10	-0.5	0.2	1

Exercise 30 *a. Calculate the standard deviation of a portfolio composed of 50% of stock A and 50% of stock C.*

b. Calculate the standard deviation of a portfolio composed of 20% of stock A, 60% of stock B and 20% of stock C.

c. Calculate the standard deviation of a portfolio composed of 70% of stock A, 60% of stock B and a short sale of C.

Exercise 31 *From finance.yahoo.com, find the historical price data on IBM stock over the previous ten years. Calculate the return each year, the mean return and the variance. Repeat for the stock of General Motors and Boeing. Hence find the expected return and variance of a portfolio consisting off 20% IBM, 30% General Motors and 50% Boeing.*

Chapter 4

The Efficient Frontier

To make a good choice we must first know the full range of alternatives. Once these are known it may be found that some can dismissed as poor, simply giving less of what we want and more of what we don't want. These alternative should be discarded. From what is left, the choice should be made. In finance terms, no investor wishes to bear unnecessary risk for the return the return that they are achieving. This implies being efficient and maximizing return for given risk. Given this, what remains is to choose the investment strategy that makes the best trade-off between risk and return. An investor needs to know more than just the fact that there is trade off between the two. What it is necessary to find is the relationship between risk and return as portfolio composition is changed. We already know that this relationship must depend on the variances of the asset returns and the covariance between them. The relationship that we ultimately construct is the efficient frontier. This is the set of efficient portfolios from which a choice is made.

4.1 Introduction

The investment decision involves the comparison of the returns and risks of different potential portfolios. The calculations of the previous chapter have shown how to determine the expected return on a portfolio and the variance of return. To make an informed choice of portfolio an investor needs to know the possible combinations of risk and return that can be achieved by alternative portfolios. Only with this knowledge is it possible to make an informed choice of portfolio.

The starting point for investigating the relationship between risk and return is a study of portfolios composed of just two risky assets with no short-selling. The relationship between risk and return that is constructed is termed the *portfolio frontier* and the shape of the frontier is shown to depend primarily upon the coefficient of correlation between the returns on the two assets. The concept of a *minimum variance portfolio* is introduced and the *efficient frontier* – the set of assets with maximum return for a given level of risk – is identified. The minimum variance portfolio is later shown to place a central role in the identification of efficient portfolios.

The restrictions on the number of assets and on short-selling are then relaxed in order to move the analysis closer to practical application. Permitting short-selling is shown to extend the portfolio frontier but not to alter its shape. Introducing additional risky assets generalizes the portfolio frontier into the *portfolio set*, but the idea of an efficient frontier is retained. The extensions are completed by allowing a risk-free asset, both with a single interest rate and differing interest rates for borrowing and lending.

The outcome of this analysis is the identification of the set of portfolios from which an investor should choose, and the set of portfolios that should not be chosen. This information is carried into the next chapter where the efficient set is confronted with preferences.

4.2 Two-Asset Portfolios

The analysis begins by considering the risk and return combinations offered by portfolios composed of two risky assets. We start by assuming that there is no risk-free asset and short sales are not possible. This simple case is the basic building block for the analysis of more general situations that relax the assumptions.

The two risky assets are labelled A and B. It is assumed that the expected return on asset A is less than that of asset B, so $\bar{r}_A < \bar{r}_B$. For any investor to choose asset A it must offer a lower variance of return than asset B. It is therefore assumed that $\sigma_A^2 < \sigma_B^2$. If these conditions were not met, either one asset would never be chosen or, if the return and variance of both were the same, the two assets would be identical and no issue of choice would arise.

A portfolio is described by proportional holdings X_A and X_B of the assets with the property that $X_A + X_B = 1$. Ruling out short sales implies that the holdings of both assets must be positive, so $X_A \geq 0$ and $X_B \geq 0$. The focus of attention is the relation between the standard deviation of the return on the portfolio, σ_p, and the expected return of the portfolio, $\bar{r}_p$, as the portfolio proportions X_A and X_B are varied. The reason for this interest is that this relationship reveals the manner in which an investor can trade risk for return by varying the composition of the portfolio.

Recall from (3.48) that the standard deviation of the return on a two-asset portfolio is given by

$$\sigma_p = \left[X_A^2\sigma_A^2 + X_B^2\sigma_B^2 + 2X_AX_B\rho_{AB}\sigma_A\sigma_B\right]^{1/2}. \tag{4.1}$$

Now consider the variances of the two assets and the proportional holdings to be given. The standard deviation of the return on the portfolio then depends only upon the value of the correlation coefficient, ρ_{AB}. This observation motivates the strategy of considering how the standard deviation/expected return relationship depends on the value of the correlation coefficient.

The analysis now considers the two limiting cases of perfect positive correlation and perfect negative correlation, followed by the intermediate case.

Case 1: $\rho_{AB} = +1$ (Perfect Positive Correlation)

The first case to consider is that of perfect positive correlation where $\rho_{AB} = +1$. As discussed in Chapter 3, this can be interpreted as the returns on the assets always rising or falling in unison.

Setting $\rho_{AB} = +1$, the standard deviation of the return on the portfolio becomes

$$\sigma_p = \left[X_A^2\sigma_A^2 + X_B^2\sigma_B^2 + 2X_AX_B\sigma_A\sigma_B\right]^{1/2}. \tag{4.2}$$

The term within the brackets is a perfect square so its square root can be written explicitly. Taking the square root gives the solution for the standard deviation as

$$\sigma_p = X_A\sigma_A + X_B\sigma_B. \tag{4.3}$$

Equation (4.3) shows that the standard deviation of the return on the portfolio is a obtained as a weighted sum of the standard deviations of the returns on the individual assets, where the weights are the portfolio proportions. This result can be complemented by employing (3.12) to observe that the expected return on the portfolio is

$$\bar{r}_p = X_A\bar{r}_A + X_B\bar{r}_B, \tag{4.4}$$

so the expected return on the portfolio is also a weighted sum of the expected returns on the individual assets.

Example 54 provides an illustration of the risk/return relationship that is described by equations (4.3) and (4.4).

Example 54 *Let asset A have expected return $\bar{r}_A = 1$ and standard deviation $\sigma_A = 2$ and asset B have expected return $\bar{r}_B = 10$ and standard deviation $\sigma_B = 8$. Table 4.1 gives the expected return and standard deviation for various portfolios of the two assets when the returns are perfectly positively correlated. These values are graphed in Figure 4.1.*

X_A	0	0.25	0.5	0.75	1
X_B	1	0.75	0.5	0.25	0
$\bar{r}_p$	10	7.75	5.5	3.25	1
σ_p	8	6.5	5	3.5	2

Table 4.1: Perfect Positive Correlation

As Example 54 illustrates, because the equations for portfolio expected return and standard deviation are both linear the relationship between σ_p and $\bar{r}_p$ is also linear. This produces a straight line graph when expected return is plotted against standard deviation. The equation of this graph can be derived

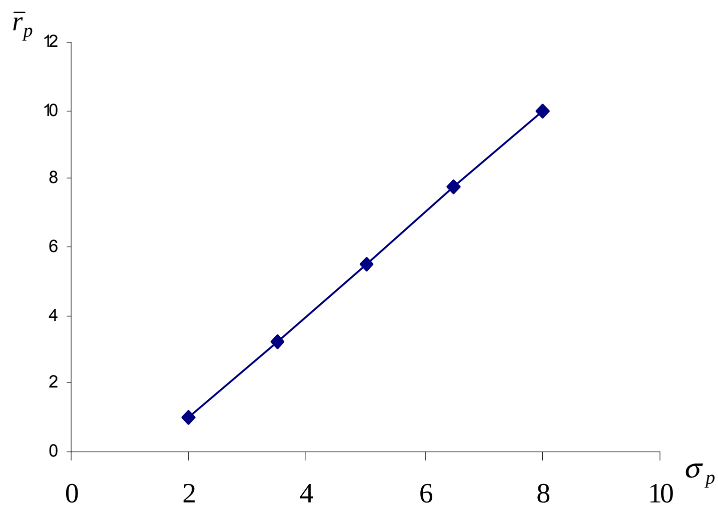


Figure 4.1: Risk and Return

as follows. The portfolio weights must sum to 1 so $X_B = 1 - X_A$. Substituting for X_B in (4.3) and (4.4), and then eliminating X_A between the equations gives

$$\bar{r}_p = \left[\frac{\bar{r}_B \sigma_A - \bar{r}_A \sigma_B}{\sigma_A - \sigma_B} \right] + \left[\frac{\bar{r}_A - \bar{r}_B}{\sigma_A - \sigma_B} \right] \sigma_p. \tag{4.5}$$

This result makes precise the details of the linear relationship between expected return and standard deviation. It can be easily checked that the data in Table 4.1 satisfy equation (4.5).

The investment implication of the fact that the frontier is a straight line is that the investor can trade risk for return at a constant rate. Therefore, when the returns on the assets are perfectly positively correlated, each extra unit of standard deviation that the investor accepts has the same reward in terms of additional expected return.

The relationship that we have derived between the standard deviation and the expected return is called the *portfolio frontier*. It displays the trade-off that an investor faces between risk and return as they change the proportions of assets A and B in their portfolio. Figure 4.2 displays the location on this frontier of some alternative portfolio proportions of the two assets. It can be seen in Figure 4.2 that as the proportion of asset B (the asset with the higher standard deviation) is increased the location moves up along the frontier. It is important to be able to locate different portfolio compositions on the frontier as this is the basis for understanding the consequences of changing the structure of the portfolio.

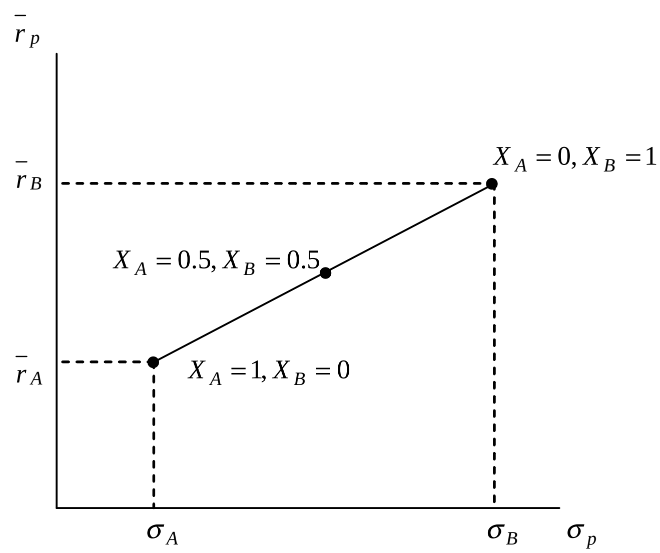


Figure 4.2: Asset Proportions on the Frontier

Case 2: $\rho_{AB} = -1$ (Perfect Negative Correlation)

The second case to consider is that of perfect negative correlation with $\rho_{AB} = -1$. Perfect negative correlation occurs when an increase in the return on one asset is met with by a reduction in the return on the other asset.

With $\rho_{AB} = -1$ the standard deviation of the portfolio becomes

$$\sigma_p = \left[X_A^2\sigma_A^2 + X_B^2\sigma_B^2 - 2X_AX_B\sigma_A\sigma_B\right]^{1/2}. \tag{4.6}$$

The term expression within the brackets is again a perfect square but this time the square root has two equally valid solutions. The first solution is given by

$$\sigma_p = X_A\sigma_A - X_B\sigma_B, \tag{4.7}$$

and the second is

$$\sigma_p = -X_A\sigma_A + X_B\sigma_B. \tag{4.8}$$

It is easily checked that these are both solutions by squaring them and recovering the term in brackets.

The fact that there are two potential solutions makes it necessary to determine which is applicable. This question is resolved by utilizing the fact that a standard deviation can never be negative. The condition that σ_p must be non-negative determines which solution applies for particular values of X_A and X_B, since when one gives a negative value for the standard deviation, the other will give a positive value. For instance, if $\sigma_B > \sigma_A$, then (4.7) will hold when X_A is large relative to X_B and (4.8) will hold when X_A is small relative to X_B.

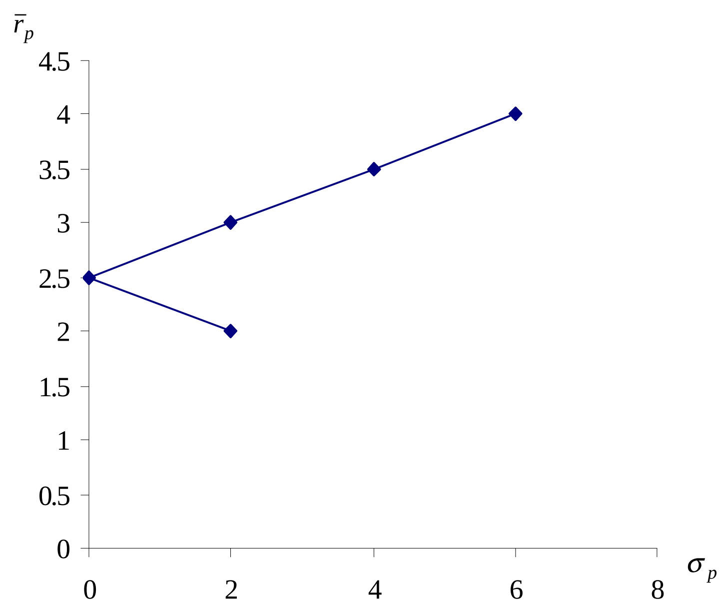


Figure 4.3: Perfect Negative Correlation

Example 55 *Let asset A have expected return $\bar{r}_A = 2$ and standard deviation $\sigma_A = 2$ and asset B have expected return $\bar{r}_B = 4$ and standard deviation $\sigma_B = 6$. Table 4.2 gives the expected return and standard deviation predicted by (4.7) and (4.8) for various portfolios of the two assets when the returns are perfectly negatively correlated. The positive values are graphed in Figure 4.3.*

X_A	0	0.25	0.5	0.75	1
X_B	1	0.75	0.5	0.25	0
$\bar{r}_p$	4	3.5	3	2.5	2
σ_p (4.7)	-6	-4	-2	0	2
σ_p (4.8)	6	4	2	0	-2

Table 4.2: Perfect Negative Correlation

The important fact about the portfolio frontier for this example is that the portfolio $X_A = \frac{3}{4}, X_B = \frac{1}{4}$ has a standard deviation of return, σ_p, that is zero. This shows that the two risky assets have combined into a portfolio with no risk (we have already observed this possibility in Section 3.3.2). That a portfolio with standard deviation of zero can be constructed from two risky assets is a general property when there is perfect negative correlation.

To find the portfolio with a standard deviation of zero, substitute $X_B = 1 - X_A$ into either (4.7) or (4.8) and set $\sigma_p = 0$. Then both (4.7) and (4.8) provide the expression

$$X_A\sigma_A - [1 - X_A]\,\sigma_B = 0. \tag{4.9}$$

Solving this equation for the proportion of asset A in the portfolio gives

$$X_A = \frac{\sigma_B}{\sigma_A + \sigma_B}, \tag{4.10}$$

which, using the fact that the proportions must sum to 1, implies the proportion of asset B is

$$X_B = \frac{\sigma_A}{\sigma_A + \sigma_B}. \tag{4.11}$$

A portfolio with the two assets held in these proportions will have a standard deviation of $\sigma_p = 0$. The values in Example 55 can be confirmed using these solutions.

Example 56 *Let asset A have standard deviation $\sigma_A = 4$ and asset B have and standard deviation $\sigma_B = 6$. The $X_A = \frac{6}{4+6} = \frac{3}{5}$ and $X_B = \frac{4}{4+6} = \frac{2}{5}$. Hence the standard deviation is*

$$\sigma_p = \left[\frac{9}{25} \times 16 + \frac{4}{25} \times 36 - 2 \times \frac{3}{5}\frac{2}{5} \times 4 \times 6\right]^{1/2} = 0.$$

The general form of the portfolio frontier for $\rho_{AB} = -1$ is graphed in Figure 4.4 where the positive parts of the equations are plotted. This again illustrates the existence of a portfolio with a standard deviation of zero. The second important observation to be made about the figure is that for each portfolio on the downward sloping section there is a portfolio on the upward sloping section with the same standard deviation but a higher return. Those on the upward sloping section therefore dominate in terms of offering a higher return for a given amount of risk. This point will be investigated in detail later.

Case 3: $-1 < \rho_{AB} < +1$

For intermediate values of the correlation coefficient the frontier must lie between that for the two extremes of $\rho_{AB} = -1$ and $\rho_{AB} = 1$. It will have a curved shape that links the positions of the two assets.

Example 57 *Let asset A have expected return $\bar{r}_A = 2$ and standard deviation $\sigma_A = 2$ and asset B have expected return $\bar{r}_B = 8$ and standard deviation $\sigma_B = 6$. Table 4.3 gives the expected return and standard deviation for various portfolios of the two assets when $\rho_{AB} = -\frac{1}{2}$. These values are graphed in Figure 4.5.*

X_A	0	0.125	0.25	0.375	0.5	0.625	0.75	0.875	1
X_B	1	0.875	0.75	0.625	0.5	0.375	0.25	0.125	0
$\bar{r}_p$	8	7.25	6.5	5.75	5	4.25	3.5	2.75	2
σ_p	6	5.13	4.27	3.44	2.65	1.95	1.50	1.52	2

Table 4.3: Return and Standard Deviation with $\rho_{AB} = -\frac{1}{2}$

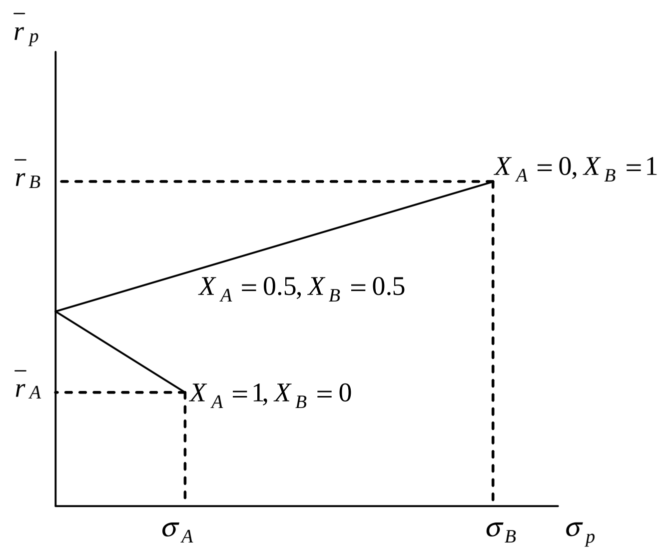


Figure 4.4: Portfolio Frontier with Perfect Negative Correlation

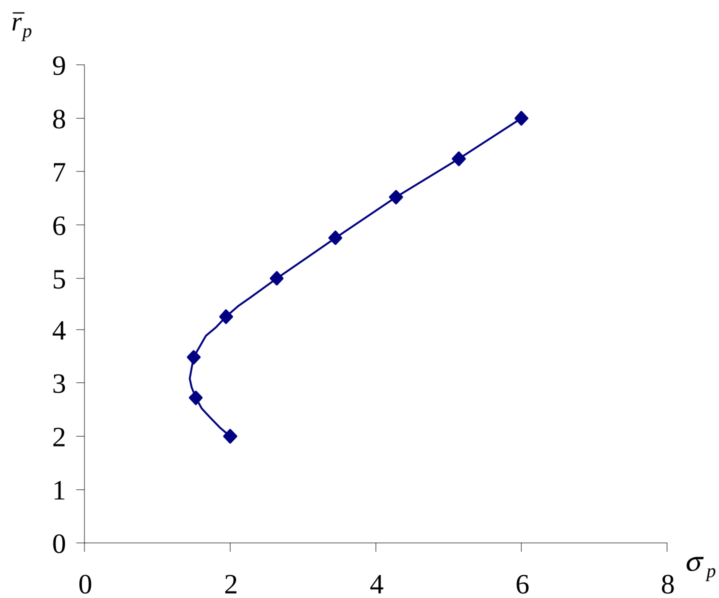


Figure 4.5: Portfolio Frontier with Negative Correlation

It can be seen in Figure 4.5 that there is no portfolio with a standard deviation of zero, but there is a portfolio that minimizes the standard deviation. This is termed the *minimum variance portfolio* and is the portfolio located at the point furthest to the left on the portfolio frontier. The composition of the minimum variance portfolio is implicitly defined by its location on the frontier. Referring back to Table 4.3 it can be seen that for the data in Example 4.5 this portfolio has a value of X_A somewhere between 0.625 and 0.875. We will see later how to calculate exactly the composition of this portfolio.

The observation that there is a minimum variance portfolio is an important one for investment analysis. It can be seen in Figure 4.5 that portfolios with a lower expected return than the minimum variance portfolio are all located on the downward-sloping section of the portfolio frontier. As was the case for perfect negative correlation, for each portfolio on the downward sloping section there is a portfolio on the upward-sloping section with a higher excepted return but the same standard deviation. Conversely, all portfolios with a higher expected return than the minimum variance portfolio are located on the upward sloping section of the frontier. This leads to the simple rule that every efficient portfolio has an expected return at least as large as the minimum variance portfolio.

Example 58 *Over the period September 1998 to September 2003, the annual returns on the stock of African Gold (traded in the UK) and Walmart (traded in the US) had a covariance of* -0.053 *(ignoring currency variations). The variance of the return on African Gold stock was* 0.047 *and that on Walmart was* 0.081*. These imply that the correlation coefficient is* $= -0.858$*. The portfolio frontier for these stocks is graphed in Figure 4.6 where point A corresponds to a portfolio composed only of African Gold stock and point B a portfolio entirely of Walmart stock.*

The analysis of the different values of the correlation coefficient in Cases 1 to 3 can now be summarized. With perfect positive correlation the portfolio frontier is upward sloping and describes a linear trade-off of risk for return. At the opposite extreme of perfect negative correlation, the frontier has a downward-sloping section and an upward-sloping section which meet at a portfolio with minimum variance. For any portfolio on the downward-sloping section there is a portfolio on the upward-sloping section with the same standard deviation but a higher return. Intermediate values of the correlation coefficient produce a frontier that lies between these extremes. For all the intermediate values, the frontier has a smoothly-rounded concave shape. The minimum variance portfolio separates inefficient portfolios from efficient portfolios. This information is summarized in Figure **??**.

The following sections are devoted to generalizing the assumptions under which the portfolio frontier has been constructed. The first step is to permit short selling of the assets but to retain all the other assumptions. The number of assets that can be held in the portfolio is then increased. Finally, a risk-free asset is introduced.

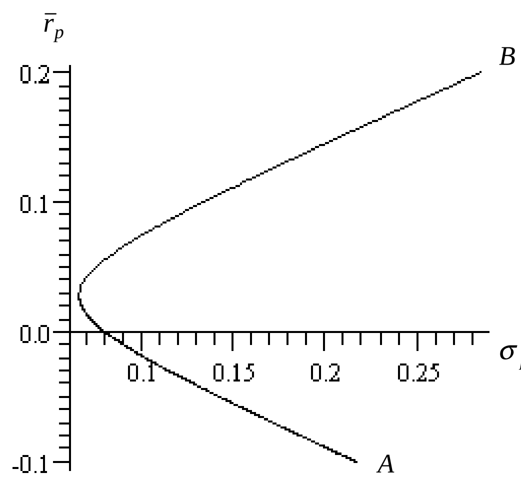


Figure 4.6: African Gold and Walmart

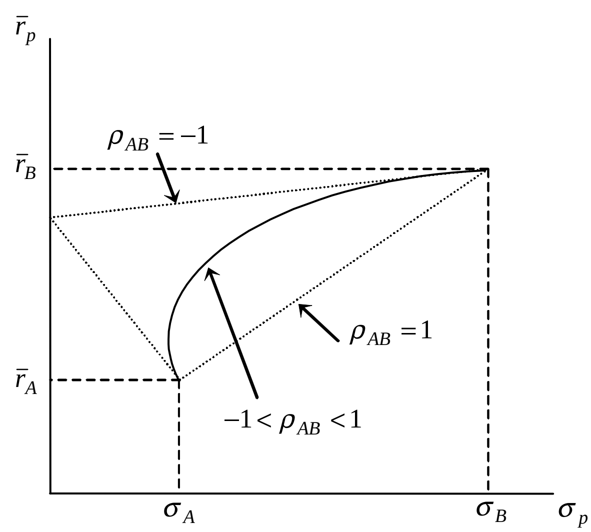


Figure 4.7: Correlation and Portfolio Frontier

4.3 Short Sales

Permitting short sales removes the non-negativity restriction on the proportions of the two assets in the portfolio. With short-selling the proportion of an asset held can be negative but the proportions must still sum to unity. This allows both positive and negative values of the portfolio proportions X_A and X_B. The only restrictions is that $X_A + X_B = 1$. For example, if asset A is sold short, so $X_A < 0$, then there must be a correspondingly long position in asset B with $X_B > 1$.

The effect of allowing short sales is to extend the frontier beyond the limits defined by the portfolios $\{X_A = 0, X_B = 1\}$ and $\{X_A = 1, X_B = 0\}$. The consequences of this change can be easily illustrated for the case of perfect positive correlation. Using (4.4), and the substitution $X_B = 1 - X_A$, the expected return is given by

$$\bar{r}_p = X_A \bar{r}_A + [1 - X_A]\, \bar{r}_B. \tag{4.12}$$

Similarly, from (4.3) the standard deviation is

$$\sigma_p = X_A \sigma_A + [1 - X_A]\, \sigma_B. \tag{4.13}$$

Without short sales, equations (4.4) and (4.3) hold only for values of X_A that satisfy $0 \leq X_A \leq 1$. But with short selling they are defined for all values of X_A that ensure $\sigma_p \geq 0$, which is the requirement that the standard deviation must remain positive. This restriction provides a range of allowable proportions X_A that is determined by σ_A and σ_B

Asset A has expected return $\bar{r}_A = 2$ and standard deviation $\sigma_A = 4$. Asset B has expected return $\bar{r}_B = 4$ and standard deviation $\sigma_B = 10$. Then $\sigma_p \geq 0$ if $X_A \leq \frac{5}{3}$ and hence $X_B \geq -\frac{2}{3}$. The portfolio frontier is graphed in Figure 4.8. Note that the choice of $X_A = \frac{5}{3}, X_B = -\frac{2}{3}$ produces a portfolio with $\bar{r}_p = 0.5$ and $\sigma_p = 0$. Therefore, short selling can produce a safe portfolio when asset returns are perfectly positively correlated.

The effect of short-selling in the general case of $-1 < \rho < 1$ is to extend the frontier as illustrated in Figure 4.9. The interpretation of points on the portfolio frontier in terms of the assets proportions needs to be emphasized. Extending the frontier beyond the portfolio composed solely of asset A is possible by going long in asset A and short-selling B. Moving beyond the location of asset B is possible by short-selling A and going long in B. The importance of these observations will be become apparent when the choice of a portfolio by an investor is considered in Chapter 5.

4.4 Efficient Frontier

The important role of the minimum variance portfolio has already been described. Every point on the portfolio frontier with a lower expected return than the minimum variance portfolio is dominated by others which has the same standard deviation but a higher return. It is from among those assets with a higher

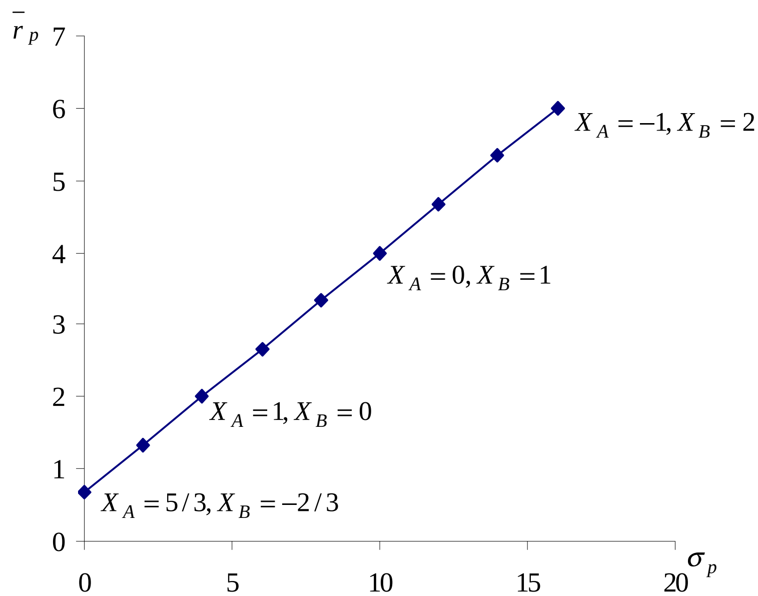


Figure 4.8: Short Selling with Perfect Positive Correlation

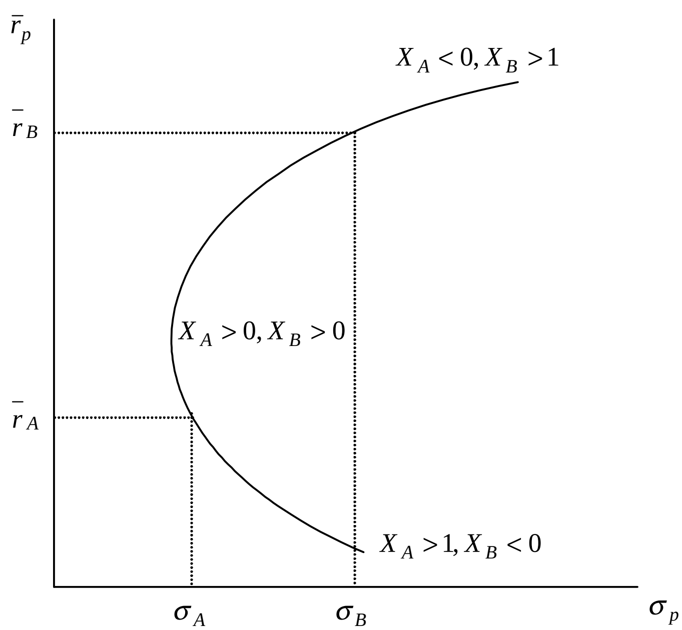


Figure 4.9: The Effect of Short Selling

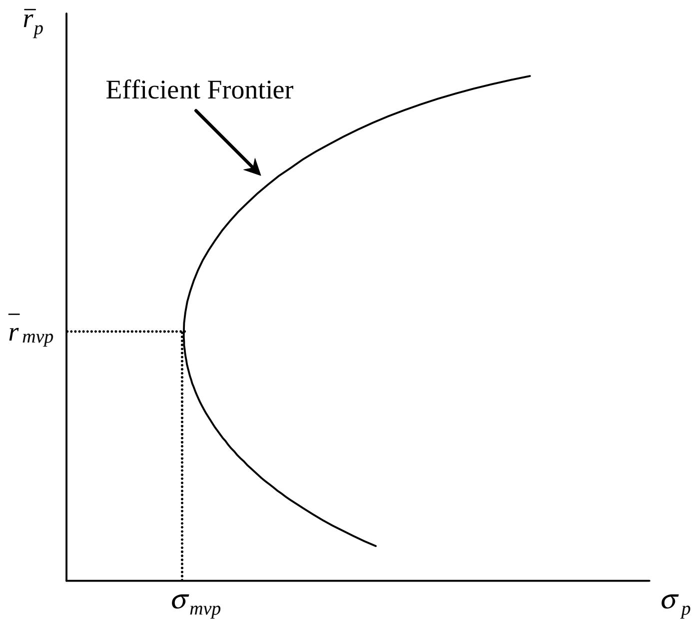


Figure 4.10: The Efficient Frontier

return than the minimum variance portfolio that an investor will ultimately make a choice. The minimum variance portfolio separates efficient portfolios that may potentially be purchased from inefficient ones that should never be purchased.

The set of portfolios with returns equal to, or higher than, the minimum variance portfolio is termed the *efficient frontier*. The efficient frontier is the upward section of the portfolio frontier and is the set from which a portfolio will actually be selected. The typical form of the efficient frontier is shown in Figure 4.10.

For every value of ρ_{AB} there is a portfolio with minimum variance. The calculation of the proportional holdings of the two assets that constitute the minimum variance portfolio is an important component of the next step in the analysis. The proportions of the two assets are found by minimizing the variance of return. The variance can in be expressed terms of the proportion of asset A alone by using the substitution $X_B = 1 - X_A$. The minimum variance portfolio then solves

$$\min_{\{X_A\}} \sigma_p^2 \equiv X_A^2 \sigma_A^2 + [1 - X_A]^2 \sigma_B^2 + 2X_A [1 - X_A] \rho_{AB} \sigma_A \sigma_B. \qquad (4.14)$$

Differentiating with respect to X_A, the first-order condition for the minimization problem is

$$\frac{\partial \sigma_p^2}{\partial X_A} \equiv X_A \sigma_A^2 - [1 - X_A] \sigma_B^2 + [1 - X_A] \rho_{AB} \sigma_A \sigma_B - X_A \rho_{AB} \sigma_A \sigma_B = 0. \qquad (4.15)$$

Solving the necessary condition for X_A gives the portfolio proportion

$$X_A = \frac{\sigma_B^2 - \sigma_A \sigma_B \rho_{AB}}{\sigma_A^2 + \sigma_B^2 - 2\sigma_A \sigma_B \rho_{AB}}. \tag{4.16}$$

For a two-asset portfolio, this portfolio proportion for asset A (and the implied proportion in asset B) characterizes the minimum variance portfolio for given values of σ_A, σ_B and ρ_{AB}.

Example 59 *With perfect positive correlation,*

$$X_A = \frac{\sigma_B^2 - \sigma_A \sigma_B}{\sigma_A^2 + \sigma_B^2 - 2\sigma_A \sigma_B} = \frac{\sigma_B}{\sigma_B - \sigma_A},$$

and with perfect negative correlation

$$X_A = \frac{\sigma_B^2 + \sigma_A \sigma_B}{\sigma_A^2 + \sigma_B^2 + 2\sigma_A \sigma_B} = \frac{\sigma_B}{\sigma_A + \sigma_B}.$$

When the assets are uncorrelated

$$X_A = \frac{\sigma_B^2}{\sigma_A^2 + \sigma_B^2}.$$

Example 60 *Using the data for Example 58, the minimum variance portfolio of African Gold stock and Walmart stock is given by*

$$\begin{aligned} X_A &= \frac{0.081 + 0.047^{\frac{1}{2}} 0.081^{\frac{1}{2}} 0.858}{0.047 + 0.081 + 2 \times 0.047^{\frac{1}{2}} 0.081^{\frac{1}{2}} 0.858} = 0.57, \\ X_B &= 0.43, \end{aligned}$$

where asset A is African Gold stock and asset B is Walmart stock. Given an expected return on African Gold stock of -0.1 and an expected return on Walmart stock of 0.2, the expected return on this portfolio is

$$\bar{r}_p = -0.1 \times 0.57 + 0.2 \times 0.43 = 0.029,$$

and the standard deviation is

$$\sigma_p = \left[0.57^2 0.047 + 0.43^2 0.081 - 2 \times 0.57 \times 0.43 \times 0.047^{\frac{1}{2}} 0.081^{\frac{1}{2}} 0.858 \right]^{\frac{1}{2}} = 0.06.$$

Refer back to Figure 4.6. In the figure point A corresponds to a portfolio composed entirely of African Gold stock and point B to a portfolio entirely composed of Walmart stock. It can be seen that the efficient frontier consists of all portfolios with a Walmart holding of at least 43% and an African Gold holding of at most 57%.

4.5 Extension to Many Assets

The next step in the analysis is to introduce additional risky assets. The first consequence of the introduction of additional assets is that it allows the formation of many more portfolios. The definition of the efficient frontier remains that of the set of portfolios with the highest return for a given standard deviation. But, rather than being found just by varying the proportions of two assets, it is now constructed by considering all possible combinations of assets and combinations of portfolios.

The process of studying these combinations of assets and portfolios is eased by making use of the following observation: a portfolio can always be treated *as if* it were a single asset with an expected return and standard deviation. Constructing a portfolio by combining two other portfolios is therefore not analytically different from combining two assets. So, when portfolios are combined, the relationship between the expected return and the standard deviation as the proportions are varied generates a curve with the form discussed above. The shape of this curve will again be dependent upon the coefficient of correlation between the returns on the portfolios.

This is illustrated in Figure 4.11 for three assets. Combining assets A and B produces the first solid curve. Combining assets C and D produces the second solid curve. Then combining portfolio 1 on first curve with portfolio 2 on second curve produces the first dashed curve. Then combining portfolio 3 on first curve with portfolio 4 on second curve produces the second dashed curve. This process can be continued by choosing a portfolio on one curve and combining it with a portfolio from another curve.

This process of forming combinations can be continued until all possible portfolios of the underlying assets have been constructed. As already described, every combination of portfolios generates a curve with the shape of a portfolio frontier. The portfolio frontier itself is the upper envelope of the curves found by combining portfolios. Graphically, it is the curve that lies outside all other frontiers and inherits the general shape of the individual curves. Hence, the portfolio frontier is always concave. The efficient frontier is still defined as the set of portfolios that have the highest return for any given standard deviation. It is that part of the portfolio frontier that begins with the minimum variance portfolio and includes all those on the portfolio frontier with return greater than or equal to that of the minimum variance portfolio. These features are illustrated in Figure 4.12.

As well as those portfolios on the frontier, there are also portfolios with return and standard deviation combinations inside the frontier. In total, the portfolio frontier and the portfolios located in the interior are called the *portfolio set*. This set is shown in Figure 4.13.

In general, the portfolio frontier is found by minimizing the standard deviation (or the variance) for a given level of return. This is analyzed in detail in the Appendix.

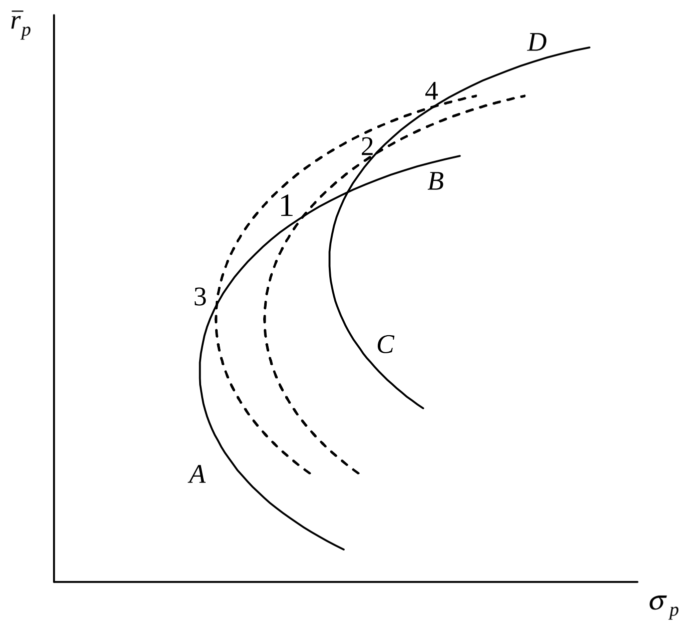


Figure 4.11: Construction of Portfolio Set

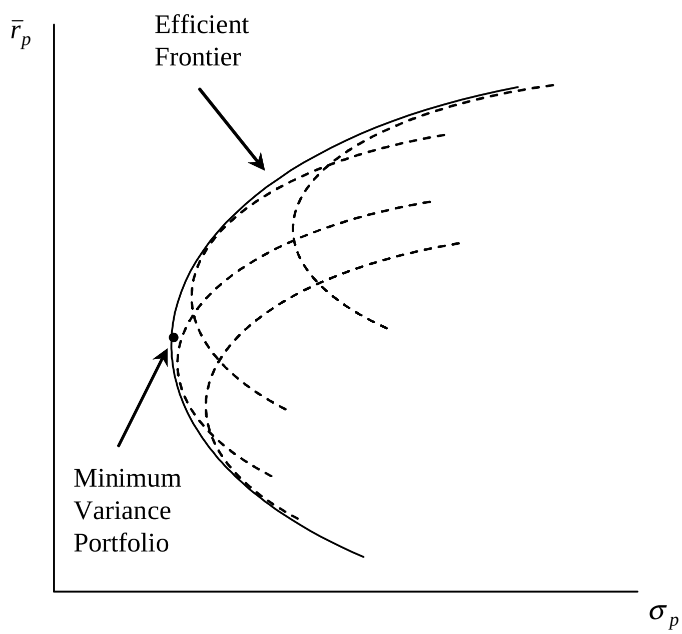


Figure 4.12: The Portfolio Frontier as an Envelope

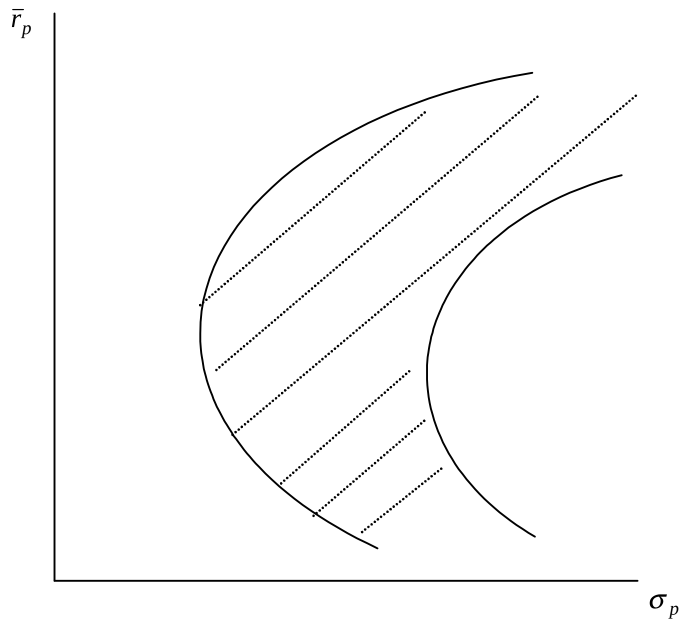


Figure 4.13: The Portfolio Set

4.6 Risk-free Asset

The previous sections have considered only risky assets. A risk-free asset is now introduced and it is shown that this has a significant effect upon the structure of the efficient frontier.

The interpretation of the risk-free asset is important for understanding the implications of the following analysis. It is usual to assume that the risk-free asset is a treasury bill issued, for instance, by the US or UK government. Investment, or going long, in the risk-free asset is then a purchase of treasury bills. The government issues treasury bills in order to borrow money, so purchasing a treasury bills is equivalent to making a short-term loan to the government. Conversely, going short in the risk-free asset means that the investor is undertaking borrowing to invest in risky assets. Given this interpretation of the risk-free asset as lending or borrowing, we can think of its return as being an interest rate.

With these interpretations, the assumption that the consumer can go long or short in a risk-free asset at a single rate of return means that the interest rate for lending is the same as that for borrowing. This is a very strong assumption that is typically at variance with the observation that the rate of interest for borrowing is greater than that for lending. We accept the assumption of the single rate in this section and relax it in the next.

An idea that we have already employed is that a portfolio of risky assets can be treated *as if* it were a single (compound) risky asset with a return and a variance. This holds as long as the proportions of the assets in the portfolio remain constant. Then combining such a portfolio with the risk-free asset is like forming a portfolio of two assets. Using this approach, it is possible to discuss the effect of combining portfolios of risky assets with the risk-free asset without needing to specify in detail the composition of the portfolio of risky assets.

Consider a given portfolio of risky assets. Denote the return on this portfolio by $\bar{r}_p$ and its variance by σ_p^2. Now consider combining this portfolio with the risk-free asset. Denote the return on the risk-free asset by r_f. Let the proportion of investment in the risky portfolio be X and the proportion in the risk-free asset be $1 - X$.

This gives an expected return on the combined portfolio of

$$\bar{r}_P = [1 - X]\, r_f + X\bar{r}_p, \tag{4.17}$$

and a standard deviation of

$$\sigma_P = \left[[1 - X]^2 \sigma_f^2 + X^2\sigma_p^2 + 2X\,[1 - X]\,\sigma_p\sigma_f\rho_{pf}\right]^{1/2}. \tag{4.18}$$

By definition the variance of the risk-free asset is zero, so $\sigma_f^2 = 0$ and $\rho_{pf} = 0$. The standard deviation of the portfolio then reduces to

$$\sigma_P = X\sigma_p. \tag{4.19}$$

Rearranging this expression

$$X = \frac{\sigma_P}{\sigma_p}. \tag{4.20}$$

Substituting into (4.17), the return on the portfolio can be expressed as

$$\bar{r}_P = \left[1 - \frac{\sigma_P}{\sigma_p}\right] r_f + \frac{\sigma_P}{\sigma_p}\bar{r}_p, \tag{4.21}$$

which can be solved for $\bar{r}_P$ to give

$$\bar{r}_P = r_f + \left[\frac{\bar{r}_p - r_f}{\sigma_p}\right]\sigma_P. \tag{4.22}$$

What the result in (4.22) shows is that when a risk-free asset is combined with a portfolio of risky assets it is possible to trade risk for return along a straight line that has intercept r_f and gradient $\frac{\bar{r}_p - r_f}{\sigma_p}$. In terms of the risk/return diagram, this line passes through the locations of the risk-free asset and the portfolio of risky assets. This is illustrated in Figure 4.14 where the portfolio p is combined with the risk-free asset.

Repeating this process for other points on the frontier gives a series of lines, one for each portfolio of risky assets. These lines have the same intercept on the vertical axis, but different gradients. This is shown in Figure 4.15 for three different portfolios 1, 2, and 3.

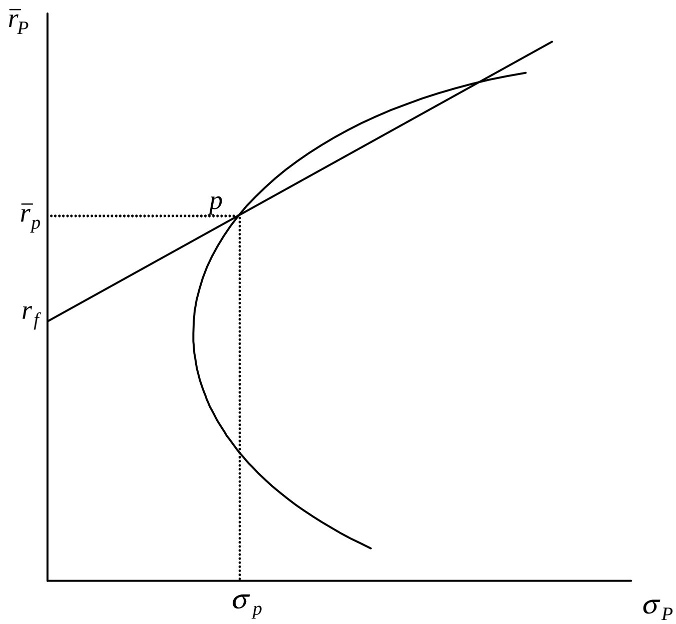


Figure 4.14: Introducing a Risk-Free Asset

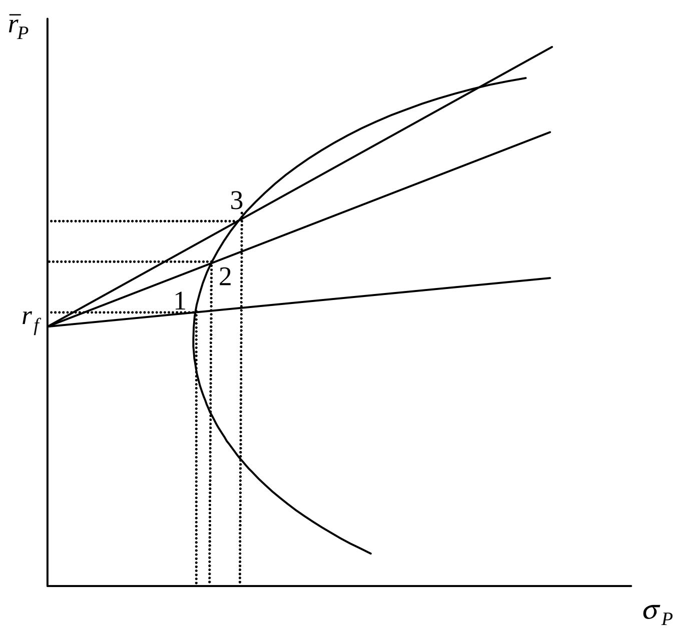


Figure 4.15: Different Portfolios of Risky Assets

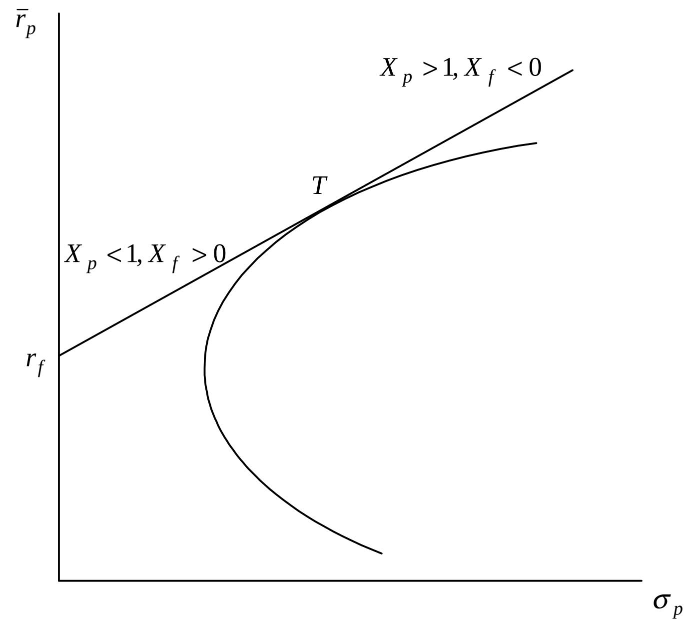


Figure 4.16: Efficient Frontier with a Risk-Free Asset

The final step in the analysis is to find the efficient frontier. Observe in Figure 4.15 that the portfolios on the line through point 3 provide a higher return for any standard deviation than those through 1 or 2. The set of efficient portfolios will then lie on the line that provides the highest return for any variance. This must be the portfolio of risky assets that generates the steepest line. Expressed differently, the efficient frontier is the line which makes the gradient $\frac{\bar{r}_p - r_f}{\sigma_p}$ as great as possible. Graphically, this line is tangential to the portfolio frontier for the risky assets. This is shown in Figure 4.16 where portfolio T is the tangency portfolio.

Consequently when there is a risk-free asset the efficient frontier is linear and all portfolios on this frontier combine the risk-free asset with alternative proportions of the tangency portfolio of risky assets. To the left of the tangency point, the investor holds a combination of the risky portfolio and the risk-free asset. To the right of the tangency point, the investor is long in the risky portfolio and short in the risk-free asset. The risky assets are always purchased in the proportions implied by the structure of the tangency portfolio. The gradient of the efficient frontier (the slope of the line) is the price of risk in terms of the extra return that has to be offered to the investor in order for them to take on additional unit of standard deviation.

Example 61 *Assume that two risky assets, A and B, are available and that their returns are uncorrelated. Letting X denote the proportion of asset A in* *the portfolio of risky assets, the tangency portfolio is defined by*

$$\max_{\{X\}} \frac{\bar{r}_p - r_f}{\sigma_p} = \frac{X\bar{r}_A + [1-X]\,\bar{r}_B - r_f}{\left[X^2\sigma_A^2 + [1-X]^2\,\sigma_B^2\right]^{\frac{1}{2}}}.$$

Differentiating with respect to X, the first-order condition is

$$\frac{\bar{r}_A - \bar{r}_B}{\left[X^2\sigma_A^2 + [1-X]^2\,\sigma_B^2\right]^{\frac{1}{2}}} - \frac{1}{2}\frac{\left[\bar{r}_A + [1-X]\,\bar{r}_B - r_f\right]\left[2X\sigma_A^2 - 2\,[1-X]\,\sigma_B^2\right]}{\left[X^2\sigma_A^2 + [1-X]^2\,\sigma_B^2\right]^{\frac{3}{2}}} = 0.$$

Solving the first-order condition gives

$$X = \frac{\sigma_B^2\,[\bar{r}_A - r_f]}{\sigma_A^2\,[\bar{r}_B - r_f] + \sigma_B^2\,[\bar{r}_A - r_f]}.$$

This analysis can be extended to consider the effect of changes in the rate of return on the risk-free asset. Assume that there are two risky assets with asset B having the higher return and standard deviation. Then as the risk-free return increases, the gradient of the efficient frontier is reduced. Moreover, the location of the tangency portfolio moves further to the right on the portfolio frontier. This increases the proportion of asset B in the risky portfolio and reduces the proportion of asset A. Through this mechanism, the rate of return on the risk-free asset affects the composition of the portfolio of risky assets.

Example 62 *Using the data for African Gold and Walmart stock in Example 58 the proportion of African Gold stock in the tangency portfolio is plotted in Figure 4.17. This graph is constructed by choosing the proportion of African Gold stock to maximize the gradient $\frac{\bar{r}_p - r_f}{\sigma_p}$ for each value of r_f. It can be seen that as the return on the risk-free asset increases, the proportion of African Gold, which has the lower return of the two assets, decreases.*

4.7 Different Borrowing and Lending Rates

It has already been noted that in practice the interest rate for lending is lower than the rate for borrowing whereas the construction of the efficient frontier in the previous section assumed that they were the same. This does not render the previous analysis redundant but rather makes it a step towards incorporating the more general situation.

Before proceeding to the analysis it is worth considering why the interest rates should be different. Fundamentally, the reason has to be the existence of some form of market inefficiency. If there were no inefficiency then all investors would be able to borrow at same rate at which they could lend. The explanation for such inefficiency can be found in the theories of information and the way in which they affect market operation. In brief, lenders are imperfectly informed

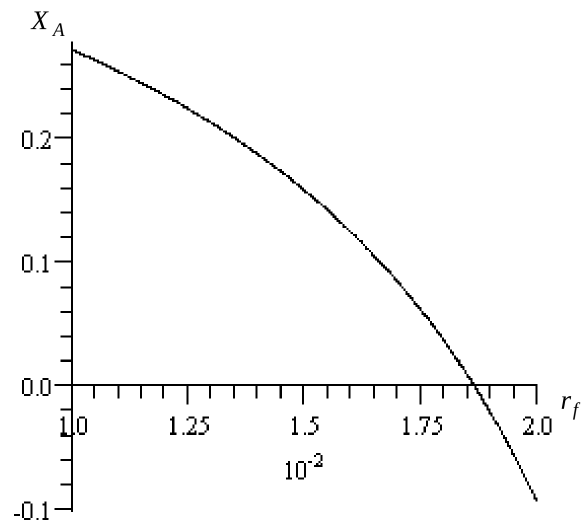


Figure 4.17: Composition of Tangency Portfolio

about the attributes of borrowers. Some borrowers, such as the US and UK government, have established strong reputations for honoring their debts and not defaulting. Consequently, they can borrow at the lowest possible rates. In contrast, private borrowers have limited reputations and lenders will be uncertain about their use of funds and consequent ability to repay. Furthermore, the borrowers are usually more informed than the lenders. These factors result in less reputable borrowers having to pay a premium on the interest rate for loans in order to compensate the lender for the increased risk.

There are two effects of there being different rates of return for lending and borrowing. Firstly, the efficient set cannot be a single line of tangency. Secondly, each investor will face an efficient set determined by the rate at which they can borrow (assuming that the lending rate corresponds to the return on treasury bills which can be purchased by any investor).

Denote the borrowing rate facing an investor by r_b and the lending rate by r_ℓ. The discussion above provides the motivation for the assumption that $r_b > r_\ell$. Denote the proportion of the investor's portfolio that is in the safe asset by X_f. If $X_f > 0$ the investor is long in the safe asset (so is lending) and earns a return r_ℓ. If $X_f < 0$ the investor is short in the safe asset (so is borrowing) and earns a return r_b. It is never rational for the investor to borrow and lend at the same time.

The structure of the efficient frontier can be developed in three steps. Firstly,

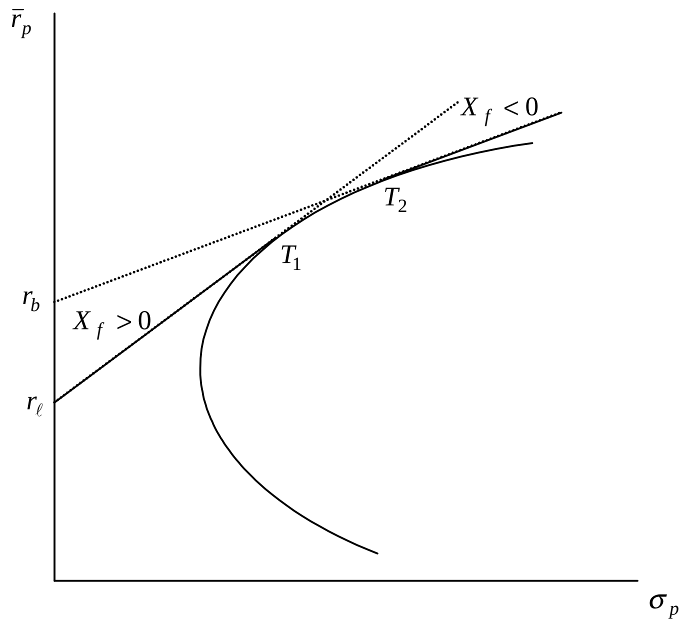


Figure 4.18: Different Returns for Borrowing and Lending

if the investor is going long in the safe asset the highest return they can achieve for a given standard deviation is found as before: the trade-off is linear and the tangency portfolio with the highest gradient is found. This gives the line in Figure 4.18 which is tangent to the portfolio frontier for the risky assets at point T_1. The difference now is that this line cannot be extended to the right of T_1: doing so would imply the ability to borrow at rate r_ℓ which we have ruled out. Secondly, if the investor borrows the efficient frontier is again a tangent line; this time with the tangency at T_2. This part of the frontier cannot be extended to the left of T_2 since this would imply the ability to lend at rate r_b. This, too, has been ruled out. Thirdly, between the tangency points T_1 and T_2, the investor is purchasing only risky assets so is neither borrowing or lending. These three sections then complete the efficient frontier.

Example 63 *If there are just two risky assets, A and B, whose returns are uncorrelated, the result in Example 61 shows that the proportion of asset A in the tangency portfolio T_1 is given by*

$$X_{A1} = \frac{\sigma_B^2 \left[\bar{r}_A - r_\ell\right]}{\sigma_A^2 \left[\bar{r}_B - r_\ell\right] + \sigma_B^2 \left[\bar{r}_A - r_\ell\right]},$$

and in the tangency portfolio T_2 by

$$X_{A2} = \frac{\sigma_B^2 \left[\bar{r}_A - r_b\right]}{\sigma_A^2 \left[\bar{r}_B - r_b\right] + \sigma_B^2 \left[\bar{r}_A - r_b\right]}.$$

It can be shown that if $\bar{r}_A < \bar{r}_B$ then $X_{A1} > X_{A2}$ so at the second tangency the proportion of the lower asset with the lower return is smaller.

In summary, when there are differing returns for borrowing and lending the efficient frontier is composed of two straight sections and one curved section. Along the first straight section the investor is long in the risk-free asset and combines this with tangency portfolio T_1. At T_1 all investment is placed in the tangency portfolio. Between T_1 and T_2 the investor purchases only risky assets with the portfolio composition changing as the move is made around this section of the portfolio frontier. Beyond T_2 the investor goes short in the risk-free asset and combines this short position with a purchase of the risky assets described by the portfolio at T_2.

4.8 Conclusions

The chapter has investigated the risk/return relationship as portfolio composition is varied. For portfolios consisting of only risky assets, a portfolio frontier is obtained whose shape depends on the correlation of asset returns. The minimum variance portfolio was defined and its role its separating efficient from inefficient portfolios was identified. From this followed the determination of the efficient frontier - the set of portfolios with return at least as great as the minimum variance portfolio. A risk-free asset was then introduced and the efficient frontier was constructed as the tangent to the portfolio set. Finally, the consequence of having different returns for borrowing and lending was considered.

The central message of this chapter is the fact that an investor is able to distinguish some portfolios which are efficient from others which are not. It is from the efficient set that a selection will ultimately be made. The second important observation is the role of the risk-free asset, and whether lending and borrowing rates are the same, in determining the structure of the efficient set. Given this characterization of the efficient set, it is now possible to move to the issue of portfolio choice.

Exercise 32 *The table provides data on the return and standard deviation for different compositions of a two-asset portfolio. Plot the data to obtain the portfolio frontier. Where is the minimum variance portfolio located?*

X	0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
$\bar{r}_p$	.08	.076	.072	.068	.064	.060	.056	.052	.048	.044	.04
σ_p	.5	.44	.38	.33	.29	.26	.24	.25	.27	.30	.35

Exercise 33 *Assuming that the returns are uncorrelated, plot the portfolio frontier without short sales when the two available assets have expected returns* 2 *and* 5 *and variances* 9 *and* 25.

Exercise 34 *Using 10 years of data from Yahoo, construct the portfolio frontier without short selling for Intel and Dell stock.*

Exercise 35 *Confirm that (4.7) and (4.8) are both solutions for the standard deviation when* $\rho_{AB} = -1$.

Exercise 36 *Given the standard deviations of two assets, what is smallest value of the correlation coefficient for which the portfolio frontier bends backward? (Hint: assuming asset A has the lower return, find the gradient of the frontier at* $X_A = 1$*.)*

Exercise 37 *Discuss the consequence of taking into account the fact that the two stocks in Example 58 are traded in different currencies. Furthermore, what role may the short data series play in this example?*

Exercise 38 *Allowing short selling, show that the minimum variance portfolios for* $\rho_{AB} = +1$ *and* $\rho_{AB} = -1$ *have a standard deviation of zero. For the case of a zero correlation coefficient, show that it must have a strictly positive variance.*

Exercise 39 *Using the data in Exercise 33, extend the portfolio frontier to incorporate short selling.*

Exercise 40 *Calculate the minimum variance portfolio for the data in Example 57. Which asset will never be sold short by an efficient investor?*

Exercise 41 *Using (4.16), explain how the composition of the minimum variance portfolio changes as the variance of the individual assets is changed and the covariance between the returns is changed.*

Exercise 42 *Calculate the minimum variance portfolio for Intel and Disney stock.*

Exercise 43 *For a two-asset portfolio, use (4.22) to express the risk and return in terms of the portfolio proportions. Assuming that the assets have expected returns of 4 and 7, variances of 9 and 25 and a covariance of* -12*, graph the gradient of the risk/return trade-off as a function of the proportion held of the asset with lower return. Hence identify the tangency portfolio and the efficient frontier.*

Exercise 44 *Taking the result in Example 61, show the effect on the tangency portfolio of (a) an increase in the return on the risk-free asset and (b) an increase in the riskiness of asset A. Explain your findings.*

Exercise 45 *What is the outcome if a risk-free asset is combined with (a) two assets whose returns are perfectly negatively correlated and (b) two assets whose returns are perfectly positively correlated?*

Exercise 46 *Prove the assertion in Example 63 that if* $\bar{r}_A < \bar{r}_B$ *then* $X_{A1} > X_{A2}$.

Chapter 5

Portfolio Selection

Choice is everything! But even when we have determined the available options, it is necessary to know exactly what we want in order to make the best use of our choices. It is most likely that we have only a vague notion of what our preferences are and how we should respond to risk. Don't immediately know this but must work from a basic feeling to clearer ideas. Consequently, want to summarize and construct preferences. We end up suggesting how people should behave. Even though some may not act this way it would be in their interests to do so.

5.1 Introduction

The process of choice involves two steps. The first step is the identification of the set of alternatives from which a choice can be made. The second step is to use preferences to select the best choice. The application of this process to investments leads to the famous *Markovitz model* of portfolio selection.

The first step of the process has already been undertaken. The efficient frontier of Chapter 4 identifies the set of portfolios from which a choice will be made. Any portfolio not on this frontier is inefficient and should not be chosen. Confronting the efficient frontier with the investor's preferences then determines which portfolio is chosen. This combination of the efficient frontier and preferences defined over portfolios on this frontier is the Markovitz model. This model is at the heart of investment theory.

The study of choice requires the introduction of preferences. The form that an investor's preferences taken when confronted with the inherent risk involved in portfolio choice is developed from a formalized description of the decision problem. This study of preferences when the outcome of choice is risky leads to the *expected utility theorem* that describes how a rational investor should approach the decision problem. Once preferences have been constructed they can be combined with the efficient frontier to solve the investor's portfolio selection problem.

There are parts of this chapter that are abstract in nature and may seem far removed from practical investment decisions. The best way to view these is as formalizing a method of thinking about preferences and decision making. Both these a slightly tenuous concepts and difficult to give a concrete form without proceeding through the abstraction. The result of the analysis is an understanding of the choice process that fits well with intuitive expectations of investor behaviour. Indeed, it would be a poor representation of choice if it did otherwise. But, ultimately, the Markovitz model very neatly clarifies how an investor's attitudes to risk and return affect the composition of the chosen portfolio.

A reader that is not deeply concerned with formalities can take most of the chapter on trust and go immediately to Section 5.5. This skips the justification for how we represent preferences but will show how those preferences determine choice.

5.2 Expected Utility

When a risky asset is purchased the return it will deliver over the next holding period is unknown. What is known, or can at least be assessed by an investor, are the possible values that the return can take and their chances of occurrence. This observation can be related back to the construction of expected returns in Chapter 3. The underlying risk was represented by the future states of the world and the probability assigned to the occurrence of each state. The question then arises as to what guides portfolio selection when the investment decision is made in this environment of risk.

The first step that must be taken is to provide a precise description of the decision problem in order to clarify the relevant issues. The description that we give reduces the decision problem to its simplest form by stripping it of all but the bare essentials.

Consider an investor with a given level of initial wealth. The initial wealth must be invested in a portfolio for a holding period of one unit of time. At the time the portfolio is chosen the returns on the assets over the next holding period are not known. The investor identifies the future states of the world, the return on each asset in each state of the world, and assigns a probability to the occurrence of each state. At the end of the holding period the returns of the assets are realized and the portfolio is liquidated. This determines the final level of wealth. The investor cares only about the success of the investment over the holding period, as measured by their final level of wealth, and does not look any further into the future.

This decision problem can be given the following formal statement:

- At time 0 an investment plan ϕ is chosen;
- There are n possible states of the world, $i = 1, ..., n$, at time 1;

- At time 0 the probability of state i occurring at time 1 is π_i;
- The level of wealth at time 1 in state i with investment plan ϕ is $W_i(\phi)$;
- At time 1 the state of the world is realized and final wealth determined.

Example 64 *An investor allocates their initial wealth between a safe asset and a risky asset. Each unit of the risky asset costs $10 and each unit of the safe asset $1. If state 1 occurs the value of the risky asset will be $15. If state 2 occurs the value of the risky asset will be $5. The value of a unit of the safe asset is $1 in both states. Letting ϕ_1 be the number of units of the risky asset purchased and ϕ_2 the number of units of the safe asset, the final wealth levels in the two states are*

$$W_1(\phi) = \phi_1 \times 15 + \phi_2,$$

and

$$W_2(\phi) = \phi_1 \times 5 + \phi_2.$$

The example shows how an investor can compute the wealth level in each state of the world. The investor also assigns a probability to each state, which becomes translated into a probability for the wealth levels. Hence, it state 1 occurs with probability π_1 then wealth level $W_1(\phi)$ occurs with probability π_1. We can safely assume that an investor prefers to have more wealth than less. But with the risk involved in the portfolio choice problem this is not enough to guide portfolio choice. It can be seen in the example that every choice of portfolio leads to an allocation of wealth across the two states. For example, a portfolio with a high value of ϕ_1 relative to ϕ_2 gives more wealth in state 1 and less in state 2 compared to a portfolio with a relatively low value of ϕ_1. The key step in the argument is to show how the wish for more wealth *within* a state translates into a set of preferences over allocations of wealth *across* states.

The first step is to formalize the assumption that the investor prefers more wealth to less. This formalization is achieved by assuming that the preferences of the investor over wealth levels, W, when these are known with certainty, can be represented by a utility function $U = U(W)$ and that this utility function has the property that $U'(W) > 0$. Hence, the higher is the level of wealth the higher is utility. We will consider the consequences of additional properties of the utility function in Section 5.3.

The utility function measures the payoff to the investor of having wealth W. Such a utility function can be interpreted in three different ways. First, the investor may actually operate with a utility function. For example, an investment fund may set very clear objectives that can be summarized in the form of a utility function. Second, the investor may act *as if* they were guided by a utility function. The utility function is then an abbreviated description of the principles that guide behavior and make them act as if guided by a utility function. Third, the utility function can be an analyst's summary of the preferences of the investor.

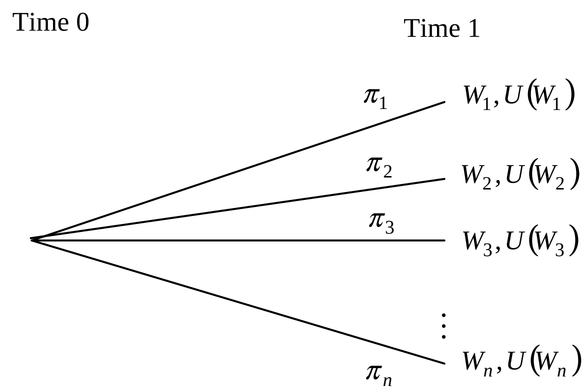


Figure 5.1: The Decision Problem

Example 65 *(i) The quadratic utility function is given by* $U = a + bW - cW^2$. *This utility function has the property that* $U'(W) > 0$ *only if* $b - 2cW > 0$. *It has the desired property only if the wealth level is not too high.*

(ii) The logarithmic utility function by $U = \log W$. *This utility function always has* $U'(W) > 0$ *provided that* $W > 0$. *This utility function is not defined for negative wealth levels (an investor in debt).*

The second step in the analysis is to impose the assumption that the investor can assess the probability of each state occurring. For state i, this probability is denoted by π_i. Because they are probabilities, it follows that $\pi_i \geq 0$, $i = 1, ..., n$, and $\sum_{i=1}^{n} \pi_i = 1$. This formulation leads to the structure shown in Figure 5.5. The interpretation of the figure is that the investor is located at time 0 looking forward to time 1. The branches emanating from time 0 are the alternative states of the world that may arise at time 1. A probability is assigned to each state. A choice of a portfolio determines the wealth level in each state. The wealth levels determine the utilities.

The difficulty facing the investor is that the choice of portfolio must be made before the state at time 1 is known. In order to analyze such *ex ante* choice in this framework a set of preferences must be constructed that incorporate the risk faced by the investor. To do this it is necessary to determine an *ex ante* evaluation of the potential income levels $\{W_1, ..., W_n\}$ that occur with probabilities $\{\pi_1, ..., \pi_n\}$ given the *ex post* preferences $U(W)$.

The preferences over wealth levels, represented by the utility function $U(W)$, can be extended to *ex ante* preferences over the random wealth levels by assuming that the investor acts consistently in such risky situations. Rationality means that the investor judges outcomes on the basis of their probabilities and payoffs, and combines multiple risky events into compound events without any inconsistencies. If the investor behaves in this way, then there preferences must satisfy the following theorem.

Theorem 1 *If a rational investor has utility of wealth* $U(W)$, *their preferences* *over risky outcomes are described by the expected utility function*

$$EU = E[U(W_i)] = \sum_{i=1}^{n} \pi_i U(W_i). \tag{5.1}$$

The theorem shows that the random consequences are evaluated by the mathematical expectation of the utility levels. This theorem has played a very important role in decision-making in risky situations because of the simplicity and precision of its conclusion. It provides the link between the evaluation of wealth when it is known with certainty and the evaluation of uncertain future wealth levels.

Example 66 *Consider an investor whose utility of wealth is represented by the utility function $U = W^{\frac{1}{2}}$. If there are three possible states of the world, the expected utility function of the investor is given by*

$$EU = \pi_1 W_1^{\frac{1}{2}} + \pi_2 W_2^{\frac{1}{2}} + \pi_3 W_3^{\frac{1}{2}}.$$

The decision of an investor is to choose a portfolio ϕ. The chosen portfolio determines a wealth level $W_i(\phi)$ in each state i. What the expected utility theorem states is that the investor should choose the portfolio ϕ to maximize expected utility subject to the cost of the portfolio being equal to the initial wealth they are investing. Let the cost of a portfolio ϕ be given by $C(\phi)$, so the investor faces the constraint be given by $W_0 = C(\phi)$. The decision problem facing the investor is then described by

$$\max_{\{a\}} E[U(W_i(\phi))] \text{ subject to } W_0 = C(\phi). \tag{5.2}$$

Example 67 *Assume an investor with an initial wealth of \$1,000 has a logarithmic utility function. Let the probability of state 1 by $\frac{2}{3}$. Assume that these is a risky asset that costs \$2 to purchase but will be worth \$3 if state 1 occurs. If state 2 occurs the risky asset will be worth \$1. Assume that there is also a risk-free asset that costs \$1 and is worth \$1 in both states. The decision problem for the investor is*

$$\max_{\{\phi_1,\phi_2\}} \frac{2}{3} \ln(3\phi_1 + \phi_2) + \frac{1}{3} \ln(\phi_1 + \phi_2),$$

subject to the budget constraint

$$1000 = 2\phi_1 + \phi_2.$$

Eliminating ϕ_2 between these equation gives

$$\max_{\{\phi_1\}} \frac{2}{3} \ln(\phi_1 + 1000) + \frac{1}{3} \ln(1000 - \phi_1).$$

differentiating with respect to ϕ_1 *the necessary condition for the maximization is*

$$\frac{2}{3(\phi_1 + 1000)} - \frac{1}{3(1000 - \phi_1)} = 0.$$

Solving the necessary condition gives

$$\phi_1 = \frac{1000}{3}.$$

This is the optimal purchase of the risky asset. The optimal purchase of the safe asset is

$$\phi_2 = 1000 - 2\phi_1 = \frac{1000}{3}.$$

This completes the general analysis of the choice of portfolio when returns are risky. The expected utility theorem provides the preferences that should guide the choice of a rational investor. The optimal portfolio then emerges as the outcome of expected utility maximization. This is a very general theory with wide applicability that can be developed much further. The following sections refine the theory to introduce more detail on attitudes to risk and how such attitudes determine the choice of portfolio.

5.3 Risk Aversion

One fundamental feature of financial markets is that investors require increased return to compensate for holding increased risk. This point has already featured prominently in the discussion. The explanation of why this is so can be found in the concept of *risk aversion.* This concept is now introduced and its relation to the utility function is derived.

An investor is described as risk averse if they prefer to avoid risk when there is no cost to doing so. A precise characterization of the wish to avoid risk can be introduced by using the idea of an *actuarially fair gamble.* An actuarially fair gamble is one with an expected monetary gain of zero. Consider entering a gamble with two outcomes. The first outcome involves winning an amount $h_1 > 0$ with probability p and the second outcome involves losing $h_2 < 0$ with probability $1 - p$. This gamble is actuarially fair if

$$ph_1 + (1 - p)h_2 = 0. \tag{5.3}$$

Example 68 *A gamble involves a probability* $\frac{1}{4}$ *of winning \$120 and a probability* $\frac{3}{4}$ *of losing \$40. The expected payoff of the gamble is*

$$\frac{1}{4} \times 120 - \frac{3}{4} \times 40 = 0.$$

If an investor is risk averse they will be either indifferent to or strictly opposed to accepting an actuarially fair gamble. If an investor is strictly risk averse then they will definitely not accept an actuarially fair gamble. Put another way, a strictly risk averse investor will never accept a gambles that does not have a strictly positive expected payoff.

Risk aversion can also be defined in terms of an investor's utility function. Let W_0 be the investor's initial wealth. The investor is risk averse if the utility of this level of wealth is higher than the expected utility arising from entering a fair gamble. Assume that $ph_1 + (1-p)h_2 = 0$, so the gamble with probabilities $\{p, 1-p\}$ and prizes $\{h_1, h_2\}$ is fair. An investor with utility function $U(W)$ is risk averse if

$$U(W_0) \geq pU(W_0 + h_1) + (1-p)U(W_0 + h_2). \tag{5.4}$$

The fact the gamble is fair allows the left-hand side of (5.4) to be written as

$$U(p(W_0 + h_1) + (1-p)(W_0 + h_2)) \geq pU(W_0 + h_1) + (1-p)U(W_0 + h_2). \tag{5.5}$$

The statement in (5.5) is just the requirement that utility function is *concave.* Strict risk aversion would imply a strict inequality in these expressions, and a strictly concave utility function.

A strictly concave function is one for which the gradient of the utility function falls as wealth increases. The gradient of the utility function, $U'(W)$, is called the marginal utility of wealth. As shown in Figure 5.2, strict concavity means that the marginal utility of wealth falls as wealth increases.

These statements can be summarized by the following:

$$\text{Risk Aversion} \Leftrightarrow U(W) \text{ concave,} \tag{5.6}$$

and

$$\text{Strict Risk Aversion} \Leftrightarrow U(W) \text{ strictly concave.} \tag{5.7}$$

Example 69 *Consider an investment for which \$10 can be gained with probability $\frac{1}{2}$ or lost with probability $\frac{1}{2}$ and an investor with initial wealth of \$100. If the investor has a logarithmic utility function then*

$$\ln(100) = 4.6052 > \frac{1}{2}\ln(100 + 10) + \frac{1}{2}\ln(100 - 10) = 4.6001.$$

This inequality shows that the logarithmic function is strictly concave so the investor is strictly risk averse.

Risk aversion is a useful concept for understanding the an investor's choice of portfolio from the efficient set. The value of the concept makes it worthwhile to review methods of measuring the degree of an investor's risk aversion. There are two alternative approaches to obtaining a measure. One methods is via the concept of a *risk premium* and the other is by defining a *coefficient of risk aversion.*

An investor's risk premium is defined as the amount that they are willing to pay to avoid a specified risk. An alternative way to express this is that the

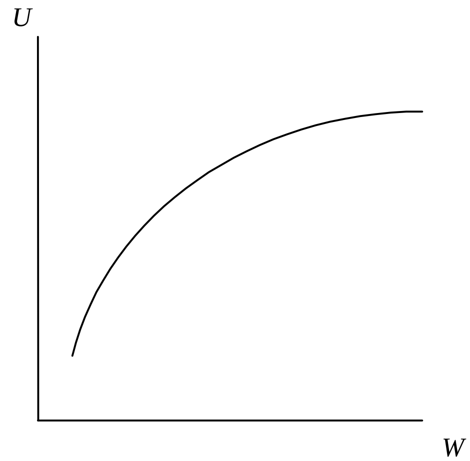


Figure 5.2: Strict Risk Aversion

risk premium is the maximum price the investor would pay for an insurance policy that completely insured the risk. The risk premium is defined relative to a particular gambles, so will vary for different gambles. But for a given gamble it can be compared across different investors to judge who will pay the lowest price to avoid risk.

Consider a gamble with two outcomes $h_1 > 0$ and $h_2 < 0$ which occur with probabilities p and $1 - p$. Assume that

$$ph_1 + [1 - p]\, h_2 \leq 0, \tag{5.8}$$

so that a risk-averse investor would prefer not to accept the gamble. The risk premium is defined as the amount the investor is willing to pay to avoid the risk. Formally, it is the amount that can be taken from initial wealth to leave the investor indifferent between the reduced level of wealth for sure and accepting the risk of the gamble. The risk premium, ρ, satisfies the identity

$$U(W_0 - \rho) = pU(W_0 + h_1) + (1 - p)U(W_0 + h_2). \tag{5.9}$$

The higher is the value of ρ for a given gamble, the more risk-averse is the investor. One way to think about this is that ρ measures the maximum price the investor is willing to pay to purchase an investment policy that ensures the gamble will be avoided.

The risk premium is illustrated in Figure 5.3. The expected utility of the gamble is $pU(W_0 + h_1) + (1 - p)U(W_0 + h_2)$, and this determines the certain income level $W_0 - \rho$ that generates the same utility. From the figure it can be seen that the more curved is the utility function, the higher is the risk premium for a given gamble. In contrast, if the utility function were linear the risk premium would be zero.

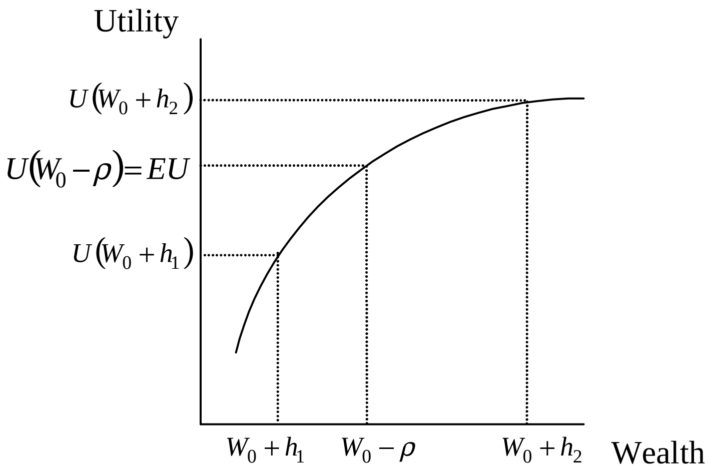


Figure 5.3: The Risk Premium

The observation that the size of the risk premium is related to the curvature of the utility function suggests the second way of measuring risk aversion. The curvature can be measured by employing the second derivative of utility. The two measures of risk aversion that are defined in this way are:

- Absolute Risk Aversion: $R_A = \frac{-U''}{U'}$;
- Relative Risk Aversion: $R_R = \frac{-WU''}{U'}$.

Absolute and relative risk aversion are equally valid as measures of risk aversion. A higher value of either measure implies a higher risk premium for any gamble. The meaning of the two measures can be investigated by considering the size of a gamble that an investor is willing to take relative to their income level. For instance, evidence indicates that investors are more willing to take a gamble of monetary value when their wealth is higher. This behavior is equivalent to absolute risk aversion being lower for investors with higher incomes. In contrast, a lower value of relative risk aversion would mean that investors with higher wealth were more likely to accept a gamble with monetary value equal to a given proportion of their wealth. There is no evidence to support this behavior.

Example 70 *For the negative exponential utility function,* $U(W) = -e^{-bW}$, *absolute risk aversion is constant with* $R_A = b$. *If an investor with this utility function and wealth* W_0 *is willing to accept a gamble with probabilities* $\{p, 1-p\}$ *and prizes* $\{h_1, h_2\}$, *the investor will accept the gamble at any wealth level.*

Example 71 *For the power utility function* $U(W) = \frac{B}{B-1} W^{\frac{B}{B-1}}, W > 0, B > 0$, *relative risk aversion is constant with* $R_R = \frac{1}{B}$. *If an investor with this* *utility function and wealth W_0 is willing to accept a gamble with probabilities $\{p, 1-p\}$ and prizes $\{W_0h_1, W_0h_2\}$, the investor will accept the gamble $\{p, 1-p\}$,$\{Wh_1, Wh_2\}$ at any wealth level W.*

5.4 Mean-Variance Preferences

The preceding sections have detailed the construction of an expected utility function that describes preferences over risky wealth levels. The key ingredients of the analysis are the set of possible wealth levels and the probabilities with which they may occur. In contrast, we have chosen to describe assets and portfolios by their returns and risks. As a consequence the preferences we are using do not sit comfortably with the characterization of portfolios. The purpose of this section is to describe the resolution of this difference.

A very important specification of expected utility for finance theory is that in which utility depends only upon the mean return and the variance of the return on a portfolio. This is important since these two characteristics of the portfolio are what underlie the concept of the efficient frontier. Preferences that depend only on the mean and variance of return can be displayed in the same diagram as the efficient frontier, and can be directly confronted with the set of efficient portfolios to investigate the selection of portfolio. The conditions under which expected utility depends on the mean and variance are now derived.

To undertake the investigation it is necessary to employ Taylor's Theorem to approximate a function. For any function the value at x_2 can be approximated by taking the value $f(x_1)$ at a different point, x_1, and adding the difference between x_1 and x_2 multiplied by the derivative of the function at x_1, so

$$f(x_2) \approx f(x_1) + f'(x_1)[x_2 - x_1]. \tag{5.10}$$

The approximation can be improved further by adding half the second derivative times the gradient squared. This process is the basis of *Taylor's Theorem* which states that for any function

$$f(x_2) = f(x_1) + f'(x_1)[x_2 - x_1] + \frac{1}{2}f''(x_1)[x_2 - x_1]^2 + R_3, \tag{5.11}$$

where R_3 is the remainder that needs to be added to make the approximation exact.

Taylor's Theorem can be applied to the utility function to determine the situations in which only the mean and variance matter. Assume that wealth random and may take any value in the range $[W_0, W_1]$. Let the expected value of wealth be $E\left[\tilde{W}\right]$. For any value of wealth $\tilde{W}$ in the range $[W_0, W_1]$ Taylor's Theorem, (5.11), can be used to write

$$\begin{aligned} U\left(\tilde{W}\right) &= U\left(E\left[\tilde{W}\right]\right) + U'\left(E\left[\tilde{W}\right]\right)\left[\tilde{W} - E\left[\tilde{W}\right]\right] \\ &\quad + \frac{1}{2}U''\left(E\left[\tilde{W}\right]\right)\left[\tilde{W} - E\left[\tilde{W}\right]\right]^2 + R_3. \end{aligned} \tag{5.12}$$

Wealth is random so the utility of wealth, $U\left(\tilde{W}\right)$, is also random. This means that the expectation of (5.12) can be taken. Two facts simplify the expectation. First, the expected deviation from the mean must satisfy $E\left[\tilde{W} - E\left[\tilde{W}\right]\right] = 0$. Second, by definition $E\left[\tilde{W} - E\left[\tilde{W}\right]\right]^2 = \sigma^2_{\tilde{W}}$. Using these facts the expected value is

$$E\left[U\left(\tilde{W}\right)\right] = U\left(E\left[\tilde{W}\right]\right) + \frac{1}{2}U''\left(E\left[\tilde{W}\right]\right)\sigma^2_{\tilde{W}} + R_3. \tag{5.13}$$

It can be seen from (5.13) that there are two sets of conditions under which only the mean and the variance of the wealth is relevant. These are either that the remainder, R_3, is exactly zero or else the remainder depends only on the mean and variance of wealth. In detail, the remainder can be written exactly as

$$R_3 = \sum_{n=3}^{\infty} \frac{1}{n!} U^{(n)}\left(E\left[\tilde{W}\right]\right)\left[\tilde{W} - E\left[\tilde{W}\right]\right]^n, \tag{5.14}$$

where $U^{(n)}$ is the n^{th} derivative of $U\left(\tilde{W}\right)$. The remainder is comprised of the additional terms that would be obtained if the approximation were continued by adding derivatives of ever higher order.

These observations are important because the mean level of wealth, $E\left[\tilde{W}\right]$, is determined by the mean return on the portfolio held by the investor. This follows since

$$E\left[\tilde{W}\right] = W_0\left(1 + \bar{r}_p\right). \tag{5.15}$$

Similarly, the variance of wealth is determined by the variance of the portfolio. Observe that

$$\begin{aligned} \sigma^2_{\tilde{W}} &= E\left[\tilde{W} - E\left[\tilde{W}\right]\right]^2 \\ &= E\left[W_0\left(1 + r_p\right) - W_0\left(1 + \bar{r}_p\right)\right]^2 \\ &= W_0^2\sigma_p^2. \end{aligned} \tag{5.16}$$

An expected utility function that depends on the mean and variance of wealth is therefore dependent on the mean and variance of the return on the portfolio.

The first situation under which only the mean and variance enter expected utility can be read directly from (5.14).

Condition 1 *If the utility function is either linear or quadratic only the mean and variance matter.*

This condition applies because if the utility function is linear or quadratic then $U^{(n)} = 0$ for any $n \geq 3$. The remainder R_3 in (5.14) is then equal to 0 whatever the values of $\left[\tilde{W} - E\left[\tilde{W}\right]\right]^n$.

If utility is quadratic, expected utility can be written as

$$\begin{aligned} E\left[U\left(\tilde{W}\right)\right] &= E\left[\tilde{W}\right] - \frac{b}{2}E\left[\tilde{W}^2\right] \\ &= E\left[\tilde{W}\right] - \frac{b}{2}\left[E\left[\tilde{W}\right]^2 + \sigma^2\left(\tilde{W}\right)\right]. \end{aligned} \tag{5.17}$$

The second situation in which only the mean and the variance enter expected utility is obtained by focusing on the terms $\left[\tilde{W} - E\left[\tilde{W}\right]\right]^n$ in the remainder. In statistical language, $\left[\tilde{W} - E\left[\tilde{W}\right]\right]^n$ is the nth central moment of the distribution of wealth. Using this terminology, the variance, $\left[\tilde{W} - E\left[\tilde{W}\right]\right]^2$, is the second central moment. It is a property of the normal distribution that the central moments, for any value of n, are determined by the value of the mean of the distribution and the variance. In short, for the normal distribution $\left[\tilde{W} - E\left[\tilde{W}\right]\right]^n = f^n\left(E\left[\tilde{W}\right], \sigma^2\left(\tilde{W}\right)\right)$ so knowing the mean and variance determines all other central moments. Therefore, for any utility function only the mean and variance matter is wealth is normally distributed.

Condition 2 *For all utility function only the mean and variance matter if wealth is distributed normally.*

If either of the conditions applies then the investor will have preferences that depend only on the mean and variance of wealth. What this means for portfolio choice is that these are the only two features of the final wealth distribution that the investor considers. The fact that the mean and variance of final wealth depend on the mean and variance of the portfolio return allows the preferences to be translated to depend only on the portfolio characteristics. Therefore, if either condition 1 or condition 2 applies, the investor has mean-variance preference that can be written as

$$U = U\left(\bar{r}_P, \sigma_P^2\right), \tag{5.18}$$

where $\bar{r}_P$ is the mean (or expected) portfolio return and σ_P^2 is its variance.

5.5 Indifference

The utility function has been introduced as a way of representing the investor's preferences over different wealth levels. Using the arguments of the previous section this can be reduced to a function that is dependent only upon the mean and variance of portfolio returns. The implications of mean-variance preferences are now developed further.

The basic concept of preference is that an investor can make a rational and consistent choice between different portfolios. An investor with mean-variance preferences makes the choice solely on the basis of the expected return and variance. This means when offered any two different portfolios the investor can provide a ranking of them using only information on the mean return and the variance of return. That is, the investor can determine that one of the two portfolios is strictly preferred to the other or that both are equally good.

The discussion of the reaction of investors to different combinations of return and risk makes it natural to assume that preferences must satisfy:

- *Non-Satiation* For a constant level of risk, more return is always strictly preferred;
- *Risk Aversion* A portfolio with higher risk can only be preferable to one with less risk if it offers a higher return.

Information about preferences can be conveniently summarized in a set of *indifference curves*. An indifference curve describes a set of portfolios which the investor feels are equally good so none of the set is preferred to any other. An indifference curve can be constructed by picking an initial portfolio. Risk is then increased slightly and the question asked of how much extra return is needed to produce a portfolio that is just as good, but no better, than the original portfolio. Conducting this test for all levels of risk then traces out a curve of risk and return combinations that is as equally good as, or *indifferent to*, the original portfolio. This curve is one indifference curve. Now consider a portfolio that has a higher return but the same risk as the original portfolio. From non-satiation, this new portfolio must be strictly better. In this case, it is said to lie on a higher indifference curve. A portfolio which is worse lies on a lower indifference curve.

The interpretation of risk aversion in terms of indifference curves is shown in Figure 5.4. Risk aversion implies that the indifference curves have to be upward sloping because more return is needed to compensate for risk. If one investor is more risk averse than another then they will require relatively more additional return as compensation for taking on an additional unit of risk. This implies that the indifference curve of the more risk averse investor through any risk and return combination is steeper than that of the less risk averse investor.

5.6 Markovitz Model

The point has now been reached at which the mean-variance preferences can be confronted with the efficient frontier. This combination is the Markovitz model of portfolio choice and is fundamental in portfolio theory. The model permits portfolio choice to be analyzed and the composition of the chosen portfolio to be related to risk aversion.

The Markovitz model makes a number of assumptions that have been implicit in the previous description but now need to be made explicit. These assumptions are:

- There are no transaction cost;
- All assets are divisible;

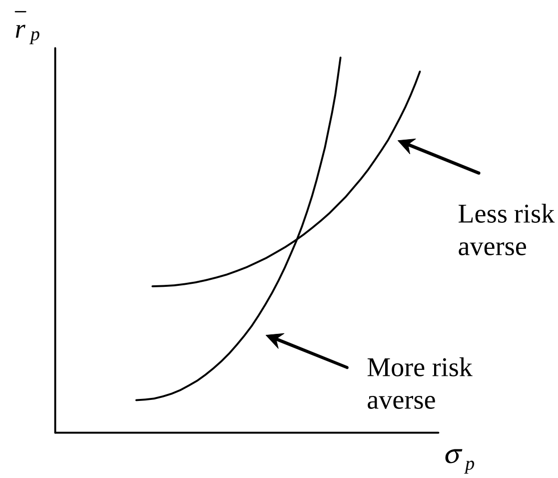


Figure 5.4: Risk Aversion and Indifference Curves

- Short selling is permitted.

The first assumption allows investors to trade costlessly so there is no disincentive to diversify or to change portfolio when new information arrives. The second assumption permits the investor to obtain an optimal portfolio no matter how awkward are the portfolio proportions. Some assets, such as government bonds, are in large denominations and indivisible. The assumption of the model can be sustained if investors can undertake indirect investments that allow the purchase of fractions of the indivisible assets. The role of short selling in extending the portfolio frontier was made clear in the previous chapter. The strong assumption is that short selling can be undertaken without incurring transaction costs.

5.6.1 No Risk-Free

Portfolio choice is first studied under the assumption that there is no risk-free asset. In this case the efficient frontier will be a smooth curve.

The optimal portfolio is the one that maximizes the mean-variance preferences given the portfolio frontier. Maximization of utility is equivalent to choosing the portfolio that lies on the highest possible indifference curve given the constraint on risk and return combinations imposed by the efficient frontier. The point on the highest indifference curve will occur at a tangency between the indifference curve and the portfolio set. Since the investor is risk averse the indifference curves are upward sloping so the tangency point must be on the efficient frontier. This means that the portfolio chosen must have a return at least as great as the minimum variance portfolio.

Figure 5.5 shows choice of two investors with different degrees of risk aversion when there are just two risky assets available. A and B denote the locations

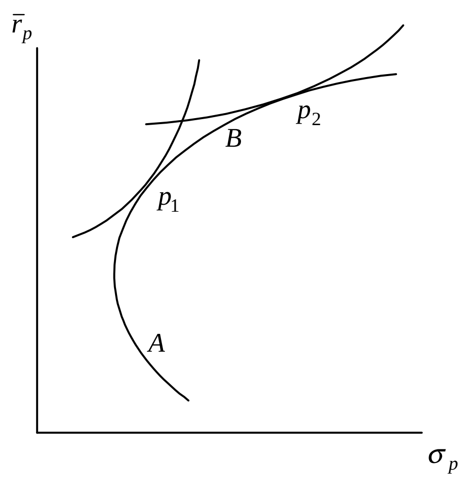


Figure 5.5: Choice and Risk Aversion

of the two available risky assets. The more risk averse investor chooses the portfolio at p_1 which combines both risky assets in positive proportions. The less risk averse investor locates at portfolio p_2. This portfolio involves going short in asset A. Since no investor chooses a portfolio with a lower return than the minimum variance portfolio, asset B will never be short-sold. In addition, any portfolio chosen must have a proportion of asset B at least as great as the proportion in the minimum variance portfolio. As risk aversion falls, the proportion of asset B increase and that of asset A falls.

The same logic applies when there are many risky assets. The investor is faced with the portfolio set and chooses a point on the upward sloping part of the frontier. The less risk-averse is the investor, the further along the upward-sloping part of the frontier is the chosen portfolio.

5.6.2 Risk-Free Asset

The introduction of a risk-free asset has been shown to have a significant impact upon efficient frontier. With the risk-free this becomes a straight line tangent to the portfolio set for the risky assets. The availability of a risk-free asset has equally strong implications for portfolio choice and leads into a mutual fund theorem.

The portfolio frontier with a risk-free asset is illustrated in Figure 5.6 with the tangency portfolio denoted by point T. The more risk-averse of the two investors illustrated chooses the portfolio p_1. This combines positive proportions of the risk-free asset and the tangency portfolio. In contrast, the less risk-averse investor chooses portfolio p_2 which involves going short in the risk-free to finance purchases of the tangency portfolio.

The important point to note is that only one portfolio of risky assets is pur-

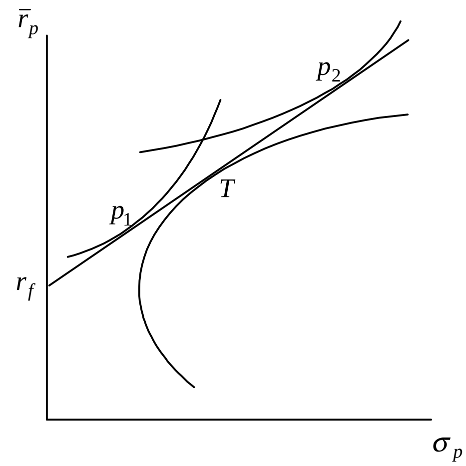


Figure 5.6: Risk-Free Asset and Choice

chased regardless of the degree of risk aversion. What changes as risk aversion changes are the relative proportions of this risky portfolio and the risk-free asset in the overall portfolio. Consequently, investors face a simple choice in this setting. They just calculate the tangency portfolio and then have to determine the mix of this with the risk-free. To do the latter, an investor just needs to evaluate their degree of risk aversion.

This observation form the basis of the *mutual fund theorem*. If there is a risk-free asset, the only risky asset that needs to be made available is a mutual fund with composition given by that of the tangency portfolio. An investor then only needs to determine what proportion of wealth should be in this mutual fund.

As a prelude to later analysis, notice that if all investors calculated the same efficient frontier then all would be buying the same tangency portfolio. As a result, this would be the only portfolio of risky assets ever observed to be purchased. There would then be no need for rigorous investment analysis since observation of other investor would reveal the optimal mix of risky assets. The assumptions necessary for this to hold and the strong implications that it has will be discussed in detail in Chapter 8.

5.6.3 Borrowing and Lending

The outcome when borrowing and lending rates are not the same is an extension of that for a single risk-free rate. In this case lending can be viewed as holding a risk-free asset with return r_ℓ and borrowing as going short in a risk-free asset with return r_b, where $r_\ell < r_b$. Two risk-free assets with different returns can co-exist since it is assumed that it is not possible to go short in the asset with return r_ℓ, nor is it possible to hold the asset with return r_b. This assumption can be justified by appealing to the existence of a market imperfection such as

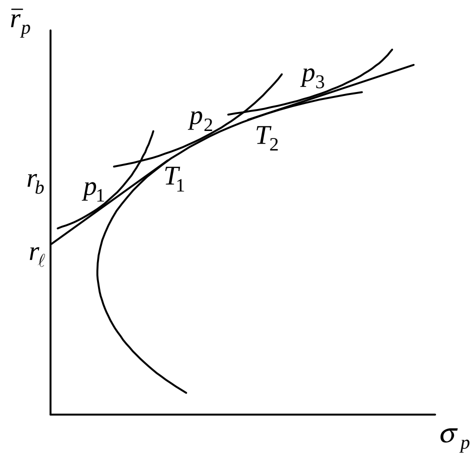


Figure 5.7: Different Interest Rates

asymmetric information between borrowers and lenders.

Figure 5.7 shows the outcome for three investors with different degrees of risk aversion. The most risk averse mixes a holding of the risk-free asset with return r_ℓ and the tangency portfolio at T_1. The less risk-averse investor purchases risky assets only, with the choice located at p_2. Finally, the investor with even less risk aversion locates at p_3 which combines the tangency portfolio T_2 with borrowing, so the investor is going short in the risk free asset with return r_b.

In this case the structure of the portfolio of risky assets held does vary as the degree of risk aversion changes. But the range of risky portfolios that will be chosen is bounded by the two endpoints T_1 and T_2. Also, the degree of risk aversion determines whether the investor is borrowing or lending.

5.7 Implications

The analysis of this chapter has several general implications for portfolio choice. Firstly, there is no simple relationship between the composition of a portfolio and risk aversion. It is always the case that an increase in risk aversion will move portfolio choice closer to the minimum variance portfolio. However, even the minimum variance portfolio may involve short-selling which is usually seen as a risky activity. This may be surprising since it is not natural to associate short-selling with what could be extreme risk aversion. Furthermore, risk-averse investors will generally bear some risk and can even bear considerable risk. The only implication of risk aversion is that an investor will not bear unnecessary risk.

To apply these methods the value of risk aversion needs to be determined. This can be done either precisely or in general terms. It can be done precisely by using experimental type approaches to test the reaction of the investor to different risky scenarios. It can be done in general terms just by discussion with the investor about their reaction to risk. Once risk aversion is known, preferences can be confronted with the efficient frontier to determine choices.

5.8 Conclusions

This chapter has introduced a formalization of the portfolio decision problem when there is uncertainty. It was shown how to model the randomness of returns via the introduction of states of nature. This model brought in preferences over wealth and lead to the expected utility theorem. The concept of risk aversion, which is a measure of reaction to risk, was then considered. The final step was to studied when utility could be reduced to mean/variance preferences.

These mean-variance preferences were then confronted with the efficient set to analyze at portfolio choice. Three different situations were considered and for each it was traced how the portfolio changed as the degree of risk aversion changed. An important observation is that when there is a single risk-free rate the investor will mix the tangency portfolio with the risk-free asset. So all that is needed is this single risky portfolio which has the form of a mutual fund.

Exercise 47 *If there are three possible future wealth levels, which occur with equal probability, and utility is given by the square root of wealth, what is the expected utility function?*

Exercise 48 *Assume there is one risky asset and one safe asset (with a return of 0) and 2 states of the world (with returns r^1 and r^2 for the risky asset) which occur with probabilities p and $1-p$. Find the optimal portfolio for an investor with the utility function $U = \frac{W^b}{b}.U = \frac{W^b}{b}$.*

Exercise 49 *Consider an investor with the utility function $U = a + bW$. Show that they will be indifferent to taking on a fair gamble. Show that if $U = a + bW^{\frac{1}{2}}$ they will not take on the fair gamble, but will if $U = a + bW^2$. Calculate the marginal utility of wealth and the degree of absolute risk aversion for each case. Comment upon the differences.*

Exercise 50 *An investor with utility function $U = \ln W$ and total wealth of $W = \$2$ is willing to enter a gamble in which \$1 can be won or lost. What must be the minimum chance of winning for the investor to participate in the gamble?*

Exercise 51 *The table provides information on portfolio returns and variances, and the satisfaction derived from several portfolios. Use this information to graph the indifference curves of the investor. Do they satisfy risk aversion? What effect does doubling the utility number attached to these curves have?*

Portfolio	1	2	3	4	5	6	7	8	9	10	11	12
$\bar{r}_p$ (%)	1	2	4	2	4	8	3	6	12	4	8	16
σ_p (%)	2	4	6	2	4	6	2	4	6	2	4	6
Utility	1	1	1	2	2	2	3	3	3	4	4	4

Exercise 52 *Consider the quadratic utility function* $U = a + bW - cW^2$. *Find the marginal utility of wealth. What happens to this as wealth increases? Does this utility function provide a good model of preferences?*

Exercise 53 *Assume there are two risky assets whose returns are uncorrelated. The expected returns of the assets are 2 and 3, and the standard deviations 5 and 6. There is also a risk-free asset with return of 1. Find the efficient frontier. When the utility function is* $U = 10\overline{r} - 0.25\left[\overline{r}^2 + \sigma_p^2\right]$, *find the optimal portfolio.*

Part III

Modelling Returns

Chapter 6

The Single Index Model

If we want to make progress it is necessary to strip away some of the details and to focus on issues of core importance. A deft application of Occam's razor will simplify the task but retain the essence. Deeper insights can be provided without losing the essentials.

6.1 Introduction

The preceeding chapters have developed a comprehensive theory of portfolio selection. This theory has some important implications for practical investment analysis especialy the identification of inefficient portfolios. We now wish to move toward application by showing how the theory can become a practical investment tool.

Given the variance-covariance matrix for the returns on a set of assets the techniques of the previous Chapter 4 can be used to calculate the efficient frontier. What this simple statement hides is the quantity of information that is needed to put this into practice for portfolios with the degree of diversification met in practice. It will be argued in this chapter that the extent of information required makes the general method impractical. What is needed is an alternative approach that can reduce the information requirement. Fortunately, such an approach is available. As a bonus the approach we describe provides an appealingly simple way of describing the riskiness of an asset.

The methods discussed in this chapter, and the next, present method for reducing the information requirement. This chapter first quantifies the extent of the information required to calculate the efficient frontier by determining the number of variances and covariances that enter into the calculations. A statistical model of asset returns designed for application to data is introduced and it is shown how this can be implemented. The implications of the model for simplifying the calculations and reducing the data requirements are then explored. Finally, the practical interpretation and application of the model is discussed. The next chapter describes a generalisation of the model.

6.2 Dimensionality

The computation of the variance of the return on a portfolio requires information on the variance of return for each of the assets in the portfolio and the covariance of the returns for each pair of assets. The computation is straightforward provided that the information is available. The difficulty in applying this method lies in obtaining the necessary information. For a portfolio of even modest size the information requirement imposes considerable demands upon the investor.

The quantity of information required can be see by returning to formula (3.49) that gave the most general version of the expression for calculalting the variance of the return on a portfolio. For a portfolio composed of N assets, the variance is calculated using the result that

$$\sigma_p^2 = \left[\sum_{i=1}^{N} \sum_{j=1}^{N} X_i X_j \sigma_{ij} \right]. \tag{6.1}$$

Observe that for each term in the first summation there are N corresponding terms in the second summation. Computation of the double sum therefore involves a total of N^2 terms, composed of the variances for each of the N assets and the $N[N-1]$ covariances.

The number of pieces of information required is actually less than N^2 because the covariance of the return on asset i with the return on asset j is equal to the covariance of the return on j with the return on i. Since $\sigma_{ij} = \sigma_{ji}$ for all i and j, not all the terms in the double summation are different. In fact, there are only $\frac{1}{2}N[N-1]$ different covariances. Adding this number of covariances to the number of variances, the total number, ν, of variances and covariances that an investor needs to know to compute the variance of the return on a portfolio of N assets is

$$\nu = N + \frac{1}{2}N[N-1] = \frac{1}{2}N[N+1]. \tag{6.2}$$

Example 72 *If a portfolio is comprised of two assets, A and B, the variance of the return can be calculated using σ_A^2, σ_B^2, and σ_{AB}. This confirms that $\nu = \frac{1}{2}2[2+1] = 3$ pieces of information are required.*

To see the implications of the formula in (6.2) for portfolios comprised of different numbers of assets consider the next example.

Example 73 *The table states the value of ν for N between 5 and 100.*

N	5	10	20	30	40	50	60	70	80	90	100
ν	15	55	210	465	820	1275	1830	2485	3240	4095	5050

To appreciate the message of this example it should be observed that a portfolio with 100 assets is not an especially diversified one. Many private investors hold portfolios with this degree of diversification and financial institutions are very likely to manage portfolios with considerably more diversification. The figures in the example show that the number of variances and covariances essentially increases at a rate proportional to the square of the number of assets. This results in the number of variances and covariances that are required rapidly becoming very large as the number of assets in the portfolio increases. The effect that this has on the information requirement can be appreciated from the next example.

Example 74 *If an institution invests in all the stocks in the S+P* 500 *index,* $125,250$ *variances and covariances need to be known to calculate the variance of the return on the portfolio.*

The implications of these observations can be explored by considering how information on variances and covariances is obtained. There are two standard sources for the information:

- Data on asset returns;
- Analysts whose job it is to follow assets.

If data is collected it can be employed to calculate variances and covariances in the way that was described in Chapter 3. The shortcoming with this approach are the demands that it places upon the data. To accurately estimate what could be several thousand variances and covariances with any degree of accuracy requires a very extensive data set. This procedure can only work if the data reflect the current situation regarding the interactions between assets. Unfortunately, if the necessary quantity of data is obtained by using information on returns stretching back into the past, then the early observations may not be representative of the current situation. The values calculated will then be poor estimates of the actual values.

The role of analysts in a financial institution is to follow a range of stocks. They attempt to develop an understanding of the management and operation of firms whose stock they follow and the industries in which the firms operate. Using this knowledge, analysts produce predictions of future returns for the stocks and an assessment of the risks. Although analysts can be employed to provide information to evaluate the variances of the returns of the stocks they follow, it is unlikely that their knowledge can contribute much to the calculation of covariances. This is a consequence of the typical structure of a brokerage firm which divides analysts into sectoral specialists. This structure is suited to inform about variances but not covariances since the links across sectors which are needed to evaluate covariances is missing.

The conclusion from this discussion is that the large numbers of variances and covariances required to evaluate the variance of return on a well-diversified portfolio cannot be computed with any reasonable degree of accuracy. This leads to a clear problem in implementing the method for constructing the efficient frontier. What is needed is an approach that simplifies the calculation.

6.3 Model and Estimation

When faced with the excessive information requirement described above the natural response is to find a means of simplifying the problem that retains its essence but removes some of the inessential detail. This is a standard modelling technique in all sciences. There is a cost to the process of simplification because a model will never explain every aspect of the data. But the cost is acceptable if the model performs sufficiently well and makes the analysis implementable.

A model is now provided that reduces the information needed to calculate the variance of the return on a portfolio. The model has the benefit of simplicity. It also provides an investor with a direct way of characterizing the riskiness of assets.

6.3.1 The Model

The starting point for the model is to ask what determines the return on an asset. Up to this point the return has just been taken as data that is entered into calculations. The perspective is now changed, and the underlying process that produces the data is considered. Consequently, the basis of the model is the specification of a process for generating the return on each asset. The process selected relates the return on every asset that is available to a single, underlying, variable. This variable is assumed to be the return on an index that summarizes the market for the set of assets that are being analyzed. The best interpretation is to view the index as a summary of financial conditions. Having a single common variable ties together the returns on different assets and by doing so simplifies the calculation of covariances.

The formal statement of the single index model is as follows. Assume there are N assets, labelled by $i = 1, ..., N$. The *single index model* assumes that the return on any asset i is determined by the process

$$r_i = \alpha_i + \beta_i r_I + \varepsilon_i, \tag{6.3}$$

where r_i is the return on asset i and r_I is the return on an index. The terms α_i and β_i are constants, and ε_i is a random error term. What this model says is that the return on the asset is linearly related to a single common influence and that this influence is summarized by the return, r_I, on an index. This return is the aggregate variable. Furthermore, the return on the asset is not completely determined by the return on the index so there is some residual variation unexplained by the index - the random error, ε_i. As will be shown below, this process for the generation of returns greatly simplifies the calculation of variance of return on a portfolio.

Before proceeding to describe the further assumptions that are made, some discussion of what is meant by the index will be helpful. The index can be an aggregate of assets such as a portfolio of stocks for all the firms in an industry or sector. Frequently the index is taken to be the market as a whole, where market here captures the idea of the set of assets that some investor might include in a portfolio. When it is, the single index model is often called the *market model* and r_I is the return on the market portfolio. As will be shown later, the market model has additional implications (that concern the average value of β_i across the assets) beyond those of the general single-index model. For the moment, attention will be focussed on the single-index model in general with the market model analyzed in Section 6.5.

Example 75 *Consider constructing a single-index model for Chevron Corp. stock. As an appropriate index for Chevron Corp. the return on all the shares in the S+P* 500 *index could be used. Doing this gives the single-index model*

$$r_C = \alpha_C + \beta_C r_{S+P} + \varepsilon_C,$$

where r_C is the return on Chevron Corp. and r_{S+P} is the return on the S+P 500*.*

The single-index model is completed by adding to the specification in (6.3) three assumptions on the structure of the errors, ε_i:

1. The expected error is zero: $E[\varepsilon_i] = 0$, $i = 1, ..., N$;
2. The error and the return on the index are uncorrelated: $E[\varepsilon_i (r_I - \bar{r}_I)] = 0$, $i = 1, ..., N$;
3. The errors are uncorrelated between assets: $E[\varepsilon_i \varepsilon_j] = 0$, $i = 1, ..., N$, $j = 1, ..., N$, $i \neq j$.

The first assumption ensures that there is no general tendency for the model to over- or under-predict the return on the asset. The second assumption ensures that the errors random and unexplained by the return on the index The third assumption requires that there is no influence other than the return on the index that systematically affects the assets. It is possible in an implementation of the model to a data set for some of these assumptions to be true and others to be false. This point is explored further below.

6.3.2 Estimation

One of the practical benefits of the single-index model is the ease with which the parameters of the model, α_i and β_i, can be estimated from historical data on returns. Reviewing this estimation method also provides further insight into the interpretation of the parameters.

The standard method for estimating the model is to collect historical data on the return on asset i and the return on the index I. The method of least squares regression is then applied to this data to estimate the parameters. Least squares regression calculates the parameters by finding the regression line which is best fit through the data points. The intercept of the regression line with the vertical axis is equal to α_i, and the gradient of the line is β_i. By best fit is meant the regression line that minimizes the sum of the squared errors, where the error

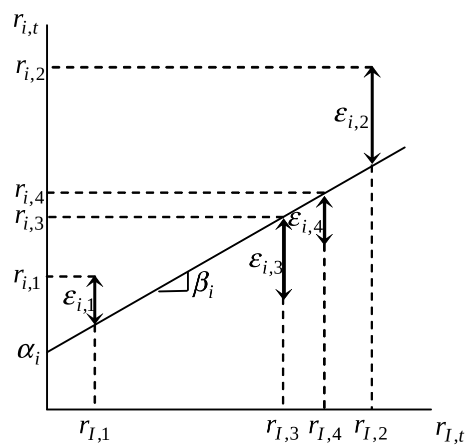


Figure 6.1: Least Squares Regression

is the vertical difference between the data point and the return predicted by the model. The values of α_i and β_i are chosen to achieve the best fit.

Figure 6.1 illustrates the method of least sqaures regression. In period t the observed return on asset i is $r_{i,t}$ and the observed return on the index is $r_{I,t}$. Four data points and the associated errors areshown in the figure. The errors are given by the vertical distances from the regression line to the data point. Some of the errors are positive, while others are negative. The regession line is adjusted by changing the intercept and the gradient until the sum of errors squared is minimized.

For given values for α_i and β_i, the error at time t is

$$e_{i,t} = r_{i,t} - (\alpha + \beta_i r_{I,t}) . \tag{6.4}$$

The data is collected for T periods so α_i and β_i are chosen to solve the following minimization

$$\min_{\{\alpha_{iI},\beta_{iI}\}} \sum_{t=1}^{T} [e_{i,t}]^2 = \sum_{t=1}^{T} [r_{i,t} - (\alpha_{iI} + \beta_{iI} r_{I,t})]^2 . \tag{6.5}$$

Differentiating the objective function with respect to α_i produces the first-order condition

$$-2 \sum_{t=1}^{T} [r_{i,t} - (\alpha_i + \beta_i r_{I,t})] = 0. \tag{6.6}$$

Differentiating the objective function with respect to β_i gives

$$-2 \sum_{t=1}^{T} r_{I,t} [r_{i,t} - (\alpha_i + \beta_i r_{I,t})] = 0. \tag{6.7}$$

Solving the pair of first-order conditions, the estimated value of β_i is given by

$$\widehat{\beta}_i = \frac{\sum_{t=1}^{T} [r_{i,t} - \bar{r}_i][r_{I,t} - \bar{r}_I]}{\sum_{t=1}^{T} [r_{I,t} - \bar{r}_I]^2}, \tag{6.8}$$

where the ˆ on β_i denotes that this is an estimated value. The formula for $\widehat{\beta}_i$ can be written more compactly as

$$\widehat{\beta}_i = \frac{\sigma_{iI}}{\sigma_I^2}, \tag{6.9}$$

so that $\widehat{\beta}_i$ is equal to the covariance of the return on the asset with the return on the index divided by the variance of the return on the index. The first-order condition (6.6) can be re-arranged and divided by T to give

$$\widehat{\alpha}_i = \bar{r}_i - \widehat{\beta}_i \bar{r}_I. \tag{6.10}$$

This completes the calculation of the least squares regression line.

Example 76 *The table provides data on the return of an asset and of an index over a five year period.*

$r_{i,t}$	*4*	*6*	*5*	*8*	*7*
$r_{I,t}$	*3*	*5*	*4*	*6*	*7*

Using this data, it can be calculated that $\bar{r}_i = 6$ *and* $\bar{r}_I = 5$. *Then*

$$\sum_{t=1}^{T} [r_{i,t} - \bar{r}_i][r_{I,t} - \bar{r}_I] = (-2)(-2) + (0)(0) + (-1)(-1) + (2)(1) + (1)(2) = 9,$$

and

$$\sum_{t=1}^{T} [r_{I,t} - \bar{r}_I]^2 = (-2)^2 + (0)^2 + (-1)^2 + (1)^2 + (2)^2 = 10.$$

These calculations give $\beta_i = \frac{9}{10}$ *and* $\alpha_i = 6 - \frac{9}{10}5 = \frac{3}{2}$. *The data points and the regression line are shown in Figure 6.2.*

The next step is to sum the samples errors defined by (6.4) to obtain

$$\sum_{t=1}^{T} e_{i,t} = \sum_{t=1}^{T} [r_{i,t} - (\alpha_{iI} + \beta_{iI} r_{I,t})]. \tag{6.11}$$

Dividing (6.11) by T gives

$$\bar{e}_i = \bar{r}_i - \widehat{\alpha}_i - \widehat{\beta}_i \bar{r}_I = 0,. \tag{6.12}$$

where the second equality is a consequence of (6.10). This result shows that the process of least squares regression ensures that the mean of the sample errors is

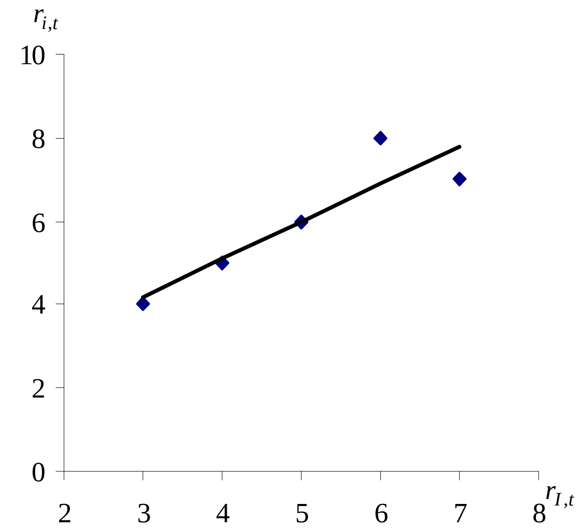


Figure 6.2: Regression Line

zero. The mean is the sample estimate of the population expectation, $E[\varepsilon_i] = 0$. Least squares regression therefore provides a sample estimate that satisfies the first assumption of the single-index model.

The first-order condition for the choice of β_i, equation (6.7), can be written as

$$\sum_{t=1}^{T} r_{I,t}\left[r_{i,t} - (\alpha_i + \beta_i r_{I,t})\right] = \sum_{t=1}^{T} r_{I,t} e_{i,t} = 0. \tag{6.13}$$

By definition

$$\begin{aligned} cov\left(e_{i,t}, r_{I,t}\right) &= \frac{1}{T}\sum_{t=1}^{T}\left(e_{i,t} - \bar{e}_i\right)\left(r_{I,t} - \bar{r}_I\right) \\ &= \frac{1}{T}\sum_{t=1}^{T} e_{i,t} r_{I,t} \\ &= 0, \end{aligned} \tag{6.14}$$

where the final equality is a consequence of (6.13). The calculated value of the covariance, $cov\left(e_{i,t}, r_{I,t}\right)$, is the sample estimate of $E\left[\varepsilon_i(r_I - \bar{r}_I)\right]$, so that least squares regression also ensures that the estimated value of the covariance satisfies the second assumption of the single-index model.

However least squares regression cannot ensure that for two assets k and j the third assumption, $E[\varepsilon_k, \varepsilon_j] = 0$, is satisfied. This is because the least squares regression considers each asset in isolation whereas the third assumption is a restriction that applies across assets.

Example 77 *Consider the result from Example 76. The estimated single-index model was*

$$r_i = \frac{3}{2} + \frac{9}{10} r_I + e_i.$$

Using this model to calculate the estimated errors provides the data in the table.

$e_{i,t}$	-0.2	0	-0.1	1.1	-0.8

It can be seen that these errors sum to zero and are uncorrelated with r_I.

The next example provides the calculation of a model for a stock with the S+P 500 index used.

Example 78 *Yahoo data.*

NOTE HERE THAT YAHOO provdies these values.

Example 79 *Yahoo data.*

6.3.3 Validity of Assumptions

Assume that an investor has collected data on the return on an index and the returns on a set of assets. The point that has to be stressed is that the single index model can always be *imposed* upon those observations. What this means is that the relation (6.3) can be always be used as a model of the process generating returns and the parameters α_i and β_i estimated for all assets. But this does not imply that it will be the *correct* model of the data: if it is not correct the estimated model will not satisfy all the assumptions of the single-index model.

Even if the assumptions are satisfied, it is not necessarily true that the model is a good one to use. As well as satisfying the assumptions it is also important to consider how much of the variation in the returns on the assets is explained by the variation in the return on the index. If it is very little, then the model is providing a poor explanation of the observations. In such a case little of the variation in the return on the asset will be explained by variation in return on the index. Instead, the non-systematic error term will be relatively large. These two points are now discussed in turn.

The estimation of the single-index model by least squares regression guarantees, by construction, that the sample mean of the estimated errors is zero for each asset and that the correlation of the error for each asset and the return on the index is also zero. Hence, for all possible observations of data, the sample equivalents of the first and second assumptions are made to hold by calculation of α_i and β_i using the least squares method. However, even if the asumptions hold this does not guarantee that the errors are small or that much of the variation in the return is explained. These points are illustrated in the following example.

Example 80 *The data on the returns on an asset,* A, *and on the returns on two indices* I_1 *and* I_2 *are given in the table.*

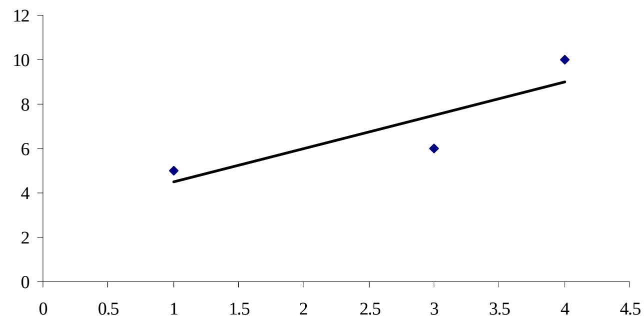


Figure 6.3: Model with Index 1

r_A	r_{I_1}	r_{I_2}
5	*1*	*4*
6	*3*	*3.75*
10	*4*	*4*

Using the data on index I_1 produces the single-index model

$$r_A = 3 + 1.5r_{I_1} + e_{AI_1},$$

which is graphed in Figure 6.3. The errors from this relationship for the three observations are $-\frac{1}{2}$, $1\frac{1}{2}$, -1, so their mean is 0. It can also be calculated that they are uncorrelated with the index. The index for this model explains 75% of the variation in the return on the asset.

Using the data on index I_2 the single-index model is

$$r_A = -16.5 + 6r_{I_2} + e_{AI_2},$$

which is graphed in Figure 6.4. The errors from this relationship are $2\frac{1}{2}$, 0, $-2\frac{1}{2}$, so their mean is 0 and they are uncorrelated with the index. The index for this model explains 10% of the variation in the return on the asset.

Both of these indices produce single index models that satisfy the assumptions on the correlation of error terms but index I_1 provides a much more informative model than index I_2.

The third assumption assumption of the single index model is that of the absence of correlation between the errors on different assets. Unlike the first two assumptions the sample equivalent of the third assumption need not hold for an application of the model estimated using least squares regression. This is just a reflection of the fact that the single index model is based on a model of how asset returns are generated and need not necessarily be true.

The failure of the third assumption to hold for a given set of data is evidence that there are other factors in addition to the index that explain the variation in asset returns. In such a case the model will need to be extended to incorporate

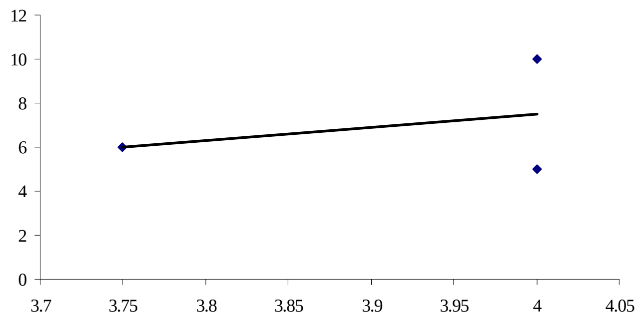


Figure 6.4: Model with Index 2

these additional correlating factors. Such extensions are the subject matter of Chapter 7.

Example 81 *Assume that the true model generating the observed returns on two assets is*

$$\begin{aligned} r_A &= 2 + 2r_{I_1} + r_{I_2} + \varepsilon_A, \\ r_B &= 3 + 3r_{I_1} + 2r_{I_2} + \varepsilon_B, \end{aligned}$$

where I_1 and I_2 are the two indices that jointly determine the asset returns. Over three periods of observation the returns and errors are

r_{I_1}	r_{I_2}	ε_A	r_A	$\varepsilon_{B,}$	r_B
1	6	0	10	1	19
2	4	1	11	$-\frac{1}{2}$	$16\frac{1}{2}$
3	1	-1	8	$-\frac{1}{2}$	$14\frac{1}{2}$

These values satisfy the requirement that $E[\varepsilon_i] = 0$ and $E[\varepsilon_A \varepsilon_B] = 0$, so the true errors are uncorrelated. If a single index model is imposed upon this data using I_1 as the index, the result would be the estimates

$$\begin{aligned} r_A &= 11\frac{2}{3} - r_{I_1} + e_A, \\ r_B &= 21\frac{1}{6} - 2\frac{1}{4}r_{I_1} + e_B, \end{aligned}$$

so the estimated errors are

e_A	$-\frac{2}{3}$	$1\frac{1}{3}$	$-\frac{2}{3}$
e_B	$\frac{1}{12}$	$-\frac{1}{6}$	$\frac{1}{12}$

These estimated errors satisfy $\bar{e}_i = 0$ and cov $(e_i, I_1) = 0$ but cov $(e_A, e_B) = -\frac{1}{9}$. The non-zero covariance of the errors is the result of imposing an incorrect model. The second index has a role to play in generating the observed returns and this is captured in the correlation of the estimated errors.

Example 82 *ANOTHER ONE FROM YAHOO: need to estiamte the model from S+P and show correlation.*

6.4 Return and Variance

The single-index model was introduced as a method for reducing the information required to calculate the variance of a portfolio. It is now shown that the single-index model achieves this aim very successfully. The application of the model to a single asset is considered first. The results derived for a single asset are then applied to a portfolio.

6.4.1 Individual Asset

The first step is to determine the implications of the process for generating returns when applied to an individual asset. The single-index model assumes that the return is determined by

$$r_i = \alpha_i + \beta_i r_I + \varepsilon_i. \tag{6.15}$$

Taking the expectation of (6.15) gives

$$\overline{r}_i = \alpha_i + \beta_i \overline{r}_I, \tag{6.16}$$

since $E[\varepsilon_i] = 0$. Hence, the expected return on the asset is determined by the expected return on the index.

Example 83 *If the expected return on an index is $\overline{r}_I = 5$, an asset described by $\alpha_i = 2$ and $\beta_i = 1.2$ has expected return*

$$\overline{r}_i = 2 + 1.2 \times 5 = 8.$$

Consider next the variance of return for an asset. The variance is defined by

$$\sigma_i^2 = E[r_i - \bar{r}_i]^2. \tag{6.17}$$

Using the single-index model to subsitute for r_i and $\bar{r}_i$ gives

$$\sigma_i^2 = E[\alpha_i + \beta_i r_I + \varepsilon_i - \alpha_i - \beta_i \bar{r}_I]^2. \tag{6.18}$$

Simplifying this expression

$$\sigma_i^2 = E[\beta_i [r_I - \bar{r}_I] + \varepsilon_i]^2. \tag{6.19}$$

Squaring the term in brackets

$$\sigma_i^2 = E\left[\beta_i^2 [r_I - \bar{r}_I]^2 + 2\varepsilon_i \beta_i [r_I - \bar{r}_I] + \varepsilon_i^2\right]. \tag{6.20}$$

The next step is to take the expectation of each of the individual terms. This gives

$$\sigma_i^2 = \beta_i^2 E[r_I - \bar{r}_I]^2 + 2\beta_i E[\varepsilon_i [r_I - \bar{r}_I]] + E\left[\varepsilon_i^2\right]. \tag{6.21}$$

The second assumption of the single-index model ensures that $E[\varepsilon_i [r_I - \bar{r}_I]] = 0$. By defintion, $E[r_I - \bar{r}_I]^2 = \sigma_I^2$ and $E\left[\varepsilon_i^2\right] = \sigma_{\varepsilon i}^2$. Using these terms the variance can finally be written as

$$\sigma_i^2 = \beta_i^2 \sigma_I^2 + \sigma_{\varepsilon i}^2. \tag{6.22}$$

Example 84 *Assume that the variance of return on the index is* $\sigma_I^2 = 16$. *For an asset with* $\beta_i = 0.8$ *and* $\sigma_{\varepsilon i}^2 = 2$, *the variance of the return on the asset is*

$$\sigma_i^2 = 0.8^2 \times 16 + 2 = 12.24.$$

From (6.22) it can be seen that the variance of the return on the asset is composed of two parts:

- Market (or systematic, or syncratic) risk, $\beta_i^2 \sigma_I^2$;
- Unique (or unsystematic, or idiosyncratic) risk, $\sigma_{\varepsilon i}^2$.

The market risk is the risk that can be predicted through knowledge of the variance of the index and the value of β_i for the asset. It is this observation that supports the interpretation of β_i as a measure of the riskiness of the asset. The unique risk of the asset is related to the asset-specific random variation that is unrelated to the index. This risk arises from all the factors that affect the return on the asset other that are not captured by the return on the index.

It should be noted that a low value of β_i does not necessarily imply a low variance of return on the asset because the idiosyncratic risk muat also be taken into account. An asset with a low value of β_i has little systematic risk, but this is only one component of total risk.

Example 85 *Assume* $\sigma_I^2 = 9$. *Then for asset A with* $\beta_{AI} = 0.9$ *and* $\sigma_{\epsilon A}^2 = 8$,

$$\sigma_A^2 = 0.9^2 \times 9 + 8 = 15.29.$$

Similarly, for asset B with $\beta_{AI} = 1.05$ *and* $\sigma_{\epsilon B}^2 = 2$,

$$\sigma_B^2 = 1.05^2 \times 9 + 2 = 11.923.$$

Asset B has a lower total variance despite having a higher value of β.

6.4.2 Portfolio Return and Variance

The expected return and the variance of the return on a portfolio can be obtained by employing similar calculations. One of the attractive features of the single index model is that the expressions for the portfolio values have the same structure as the expression for an individual asset.

Consider a portfolio of N assets, with the assets are held in proportions $X_1, ..., X_n$. The return on the portfolio is given by the weighted average of the returns on the assets

$$r_p = \sum_{i=1}^{N} X_i r_i. \tag{6.23}$$

Applying the single-index model to describe the return on each asset i determines the portfolio return

$$\begin{aligned} r_p &= \sum_{i=1}^{N} X_i \left[\alpha_i + \beta_i r_I + \varepsilon_i\right] \\ &= \sum_{i=1}^{N} X_i \alpha_i + \sum_{i=1}^{N} X_i \beta_i r_I + \sum_{i=1}^{N} X_i \varepsilon_i. \end{aligned} \tag{6.24}$$

Define the portfolio values of α and β by $\alpha_p = \sum_{i=1}^{N} X_i \alpha_i$ and $\beta_p = \sum_{i=1}^{N} X_i \beta_i$. The return on the portfolio can then be written as

$$r_p = \alpha_p + \beta_p r_I + \varepsilon_p, \tag{6.25}$$

where $\varepsilon_p = \sum_{i=1}^{N} X_i \varepsilon_i$. Observe how the single index model for each asset becomes the single index model for the portfolio. The same linear form carries over from one to the other. In addition, the error term ε_p satisfies $E\left[\varepsilon_p\right] = \sum_{i=1}^{N} X_i E\left[\varepsilon_i\right] = 0$.

Example 86 *Consider a portfolio comprised of two assets A and B with* $\alpha_A = 2$, $\beta_A = 0.6$ *and* $\alpha_B = 3$, $\beta_B = 1.2$. *Then*

$$\alpha_p = 2X_A + 3X_B,$$

and

$$\beta_p = 0.6X_A + 1.2X_B.$$

If the portfolio proportions of the two assets are $X_A = X_B = \frac{1}{2}$ *the single index model for the portfolio is*

$$r_p = 2.5 + 0.9r_I + \varepsilon_p.$$

The expected return on the portfolio is obtained by taking the expectation of (6.25). This gives

$$\begin{aligned} \bar{r}_p &= E\left[\alpha_p + \beta_p r_I + \varepsilon_p\right] \\ &= \alpha_p + \beta_p \bar{r}_I. \end{aligned} \tag{6.26}$$

The portfolio expected return is derived from the the expected return on the index and the values of α and β characterizing the portfolio.

Example 87 *Put in a portfolio from Yahoo to show* α, β. *Must give the number of units of each and the data in the table.*

Asset	α	β	*Price*

Then calculate the proportions and from this find the expected return expression on the portfolio.

The calculation of return is interesting but it is only a step toward the achievement of a simplified method of calculating the variance of return. The expression for the variance of return is now developed using the single-index model

The variance of the return on a portfolio is defined by

$$\sigma_p^2 = E\left[r_p - \bar{r}_p\right]^2. \tag{6.27}$$

Using the single index model (6.25) for the return on a portfolio and (6.26) determining the expected return the variance becomes

$$\sigma_p^2 = E\left[\alpha_p + \beta_p r_I + \varepsilon_p - \alpha_p - \beta_p \bar{r}_I\right]^2. \tag{6.28}$$

Squaring the bracketed term gives

$$\begin{aligned} \sigma_p^2 &= E\left[\beta_p^2 \left[r_I - \bar{r}_I\right]^2 + 2\beta_p \varepsilon_p \left[r_I - \bar{r}_I\right] + \varepsilon_p^2\right] \\ &= \beta_p^2 E\left[\left[r_I - \bar{r}_I\right]^2\right] + 2\beta_p E\left[\varepsilon_p \left[r_I - \bar{r}_I\right]\right] + E\left[\varepsilon_p^2\right]. \end{aligned} \tag{6.29}$$

The second assumption of the single index model implies that $E\left[\varepsilon_p \left[r_I - \bar{r}_I\right]\right] = 0$. The third assumption can be used to write

$$E\left[\varepsilon_p^2\right] = E\left[\left(\sum_{i=1}^{N} X_i \varepsilon_i\right)^2\right] = \sum_{i=1}^{N} X_i^2 E\left[\varepsilon_i^2\right]. \tag{6.30}$$

By defining $\sigma_{\varepsilon p}^2 \equiv \sum_{i=1}^{N} X_i^2 E\left[\varepsilon_i^2\right]$ the variance of the return on the portfolio can be written in compact form as

$$\sigma_p^2 = \beta_{pI}^2 \sigma_I^2 + \sigma_{\epsilon p}^2. \tag{6.31}$$

The variance of return for the portfolio mirrors that for an individual asset. There is a systematic component, $\beta_{pI}^2 \sigma_I^2$, related to the variance of return on the index and a non-systematic component, $\sigma_{\epsilon p}^2$.

An alternative expression for the variance that emphasizes the role played by the individual assets is

$$\sigma_p^2 = \left[\sum_{i=1}^{N} X_i \beta_i\right]^2 \sigma_I^2 + \left[\sum_{i=1}^{N} X_i^2 \sigma_{\varepsilon i}^2\right]. \tag{6.32}$$

In agreement with the definitions for an individual asset, the first term of this expression is the systematic variance and the second term the non-systematic variance. It also shows that holding a relatively high proportion of high-beta assets raises the variane of return on the portfolio.

Example 88 *Consider a portfolio of two assets. Asset A is described by* $\beta_A = 0.75$, *and* $\sigma_{\varepsilon A}^2 = 2$. *Asset B is described by* $\beta_B = 1.5$, *and* $\sigma_{\epsilon B}^2 = 4$. *Let the* *variance of the return on the index be $\sigma_I^2 = 25$. The variance of the return on the portfolio is*

$$\sigma_p^2 = [X_A 0.75 + X_B 1.5]^2 25 + \left[X_A^2 2 + X_B^2 4\right].$$

If $X_A = \frac{1}{3}$, then

$$\sigma_p^2 = \left[\frac{1}{3}0.75 + \frac{2}{3}1.5\right]^2 25 + \left[\frac{1}{3}^2 2 + \frac{2}{3}^2 4\right] = 41.063.$$

Expression (6.31), or equivalently (6.32), for the variance of the return of a portfolio shows very clearly the simplifying effect of imposing the single index model. To calculate the variance it is only necessary to know the values of β_i and $\sigma_{\varepsilon i}^2$ for the assets $i = 1, ..., N$ in the portfolio, and the value of σ_I^2 for the variance of return on the index. In total this requires only $2N + 1$ pieces of information. Compare this to the $\frac{1}{2}N[N+1]$ pieces of information that are required for computing the variance of return without the benefit of the single index model.

Example 89 *For a portfolio of the shares of the S+P 500 companies, the single-index model requires knowledge of 501 variances and 500 betas. This is a significant reduction from the 125,250 variances and covariances necessary without the use of the single-index model.*

These observations confirm the reduction in information requirement that is achieved when the single-index model is imposed. It must be stressed that this is not without a cost since some information is necessarily lost in using the linear relationship to model asset returns – this is explored in the examples below. But for large portfolios this is a cost that is often worth paying because it turns the theory into a practical and applicable tool.

6.4.3 Diversification

Further insight into the implications of applying the single-index model can be obtained by considering the variance of return for a well-diversified portfolio. We know already that diversification can eliminate the individual variation of the assets so that the variance of a diversified portfolio is determined by the average covariance. A similar conclusion will now be shown to be true for the single index model. In this case diversification will eliminate the non-systematic component of portfolio variance, leaving only the variance due to the index.

Consider a diversified portfolio that is evenly held, so $X_i = \frac{1}{N}$ for each of the N assets in the portfolio. Using the single-index model, the variance of return on the portfolio is the sum of systematic and non-systematic risk. Consider first the non-systematic risk on the portfolio. This is given by

$$\sigma_{\varepsilon p}^2 = \sum_{i=1}^{N} X_i^2 \sigma_{\varepsilon i}^2 = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{N} \sigma_{\varepsilon i}^2. \tag{6.33}$$

The term $\sum_{i=1}^{N} \frac{1}{N}\sigma_{\varepsilon i}^2$ is the mean non-systematic risk for the assets in the portfolio. Since this mean is finite it can be seen directly that $\sigma_{\varepsilon p}^2$ tends to 0 as N tends to infinity. Consequently, for a diversified portfolio the non-systematic risk can be diversified away. This is a consequence of the fact that this component of risk is, by definition, independent across assets and therefore cancels out in a diversified portfolio. The non-systematic risk tends to zero at a rate proportional to $1/N$, so that diversification the portfolio need not be very extreme for the non-systematic risk to become insignificant.

Example 90 *Consider three stocks from Yahoo and get the* $\sigma_{\varepsilon i}^2$ *and* σ_I^2. *Assume these are typical so get the mean* $\bar{\sigma}_{\varepsilon}^2$. *Then look at* $\left[\frac{1}{N}\right]\bar{\sigma}_{\varepsilon}^2$ *and at* $\frac{\sigma_I^2}{\left[\frac{1}{N}\right]\bar{\sigma}_{\varepsilon}^2}$ *to construct an argument about the rate at which it decays.*

The systematic component of risk for the portfolio is given by

$$\beta_p^2\sigma_I^2 = \left[\sum_{i=1}^{N} X_i\beta_i\right]^2 \sigma_I^2 = \left[\sum_{i=1}^{N} \frac{1}{N}\beta_i\right]^2 \sigma_I^2 = \overline{\beta}^2\sigma_I^2, \tag{6.34}$$

where $\overline{\beta}$ is the mean value of β_i.

Putting these observations together, the variance of the portfolio

$$\sigma_p^2 = \beta_{pI}^2\sigma_I^2 + \sigma_{\epsilon p}^2, \tag{6.35}$$

tends to the value

$$\sigma_p^2 = \overline{\beta}_I^2\sigma_I^2, \tag{6.36}$$

as N tends to infinity. For a well-diversified portfolio, only the systematic risk remains. This can be interpreted as the basic risk that underlies the variation of all assets. From this perspective, σ_I^2 can be called undiversifiable market risk and $\sigma_{\epsilon p}^2$ diversifiable risk.

The calculations have shown that if the single-index model is imposed the only data required to calculate the variance of return for a diversified portfolio is the beta for each of the assets in the portfolio and the variance of the return on the index. This has reduced the data requirement to $N+1$ pieces of information.

This argument is appealing but it must be noted that it is subject to two limitations to this argument. First, the single-index model may be a poor fit to the data, so that the index does not actually explain much of the variance of asset returns. Alternatively, the single-index model might simply not be the correct model of the data. Second, the argument that the non-systemative risk can be ignored is only really justified in the limit as the number of assets increases without bound. There is no guarantee that ignoring the non-systematic error will be a good approximation for any portfolio with limited diversification.

The idea of ignoring the non-systematic on the basis of the diversification argument has to be handled very carefully. An incorrect application can lead to mistaken conclusions. The following two examples show what happens if the non-systematic risk if ignored in the calculation of the portfolio frontier.

Example 91 *Consider a portfolio with two assets. If the idiosyncratic error is ignored the portfolio frontier for these assets is described by the pair of equations*

$$\begin{aligned}\bar{r}_p &= [X_A\alpha_A + X_B\alpha_B] + [X_A\beta_A + X_B\beta_B]\,\bar{r}_I,\\ \sigma_p &= [X_A\beta_A + X_B\beta_B]\,\sigma_I.\end{aligned}$$

But $X_B = 1 - X_A$ *so*

$$\bar{r}_p = [X_A[\alpha_A - \alpha_B] + \alpha_B] + \frac{\sigma_p}{\sigma_I}\bar{r}_I.$$

Hence, the portfolio frontier is predicted to be linear. This is false except in the special case of perfect positive correlation. Ignoring the idiosyncratic leads to a incorrect conclusion.

When there are more that two assets in the portfolio the mistakes that can be made are even greater. This is demonstrated by the next example.

Example 92 *Consider a portfolio with three assets. If the idiosyncratic error is ignored the portfolio frontier for these assets is described by the pair of equations*

$$\begin{aligned}\bar{r}_p &= [X_A\alpha_A + X_B\alpha_B + X_C\alpha_C] + [X_A\beta_A + X_B\beta_B + X_C\beta_C]\,\bar{r}_I,\\ \sigma_p &= [X_A\beta_A + X_B\beta_B + X_C\beta_C]\,\sigma_I.\end{aligned}$$

But $X_C = 1 - X_A - X_B$ *so*

$$\bar{r}_p = [X_A[\alpha_A - \alpha_C] + X_B[\alpha_B - \alpha_C] + \alpha_C] + \frac{\sigma_p}{\sigma_I}\bar{r}_I.$$

Notice that if $\alpha_A - \alpha_C \neq 0$ *then the frontier says that for any* σ_p *an infinite return can be obtained by making* X_A *large if* $\alpha_A - \alpha_C > 0$ *or by going short in* X_A *if* $\alpha_A - \alpha_C < 0$. *This will not be a correct view of the position.*

In a diversified portfolio the non-systematic risk tends to zero. This can be used as an argument for ignoring this component of risk when making a calculation of the variance of return on a portfolio. But care must be taken in the use to which such approximations are put. If the non-systematic error is ignored and then the portfolio frontier calculated a very wrong answer can be obtained. Like any approximation technique ignoring the non-systematic risk has to be usedundertaken wisely with knowledge of the limits of its applicability.

6.5 Market Model

So far there has not been much attention paid to the choice of index to use in an application of the single-index model. It was observed that the index should be some broad aggregate reflecting the market in which the risky assets comprising the portfolio are traded. For example, the S+P 500 index is representative of the market for a portfolio comprising stocks of large US corporations.

The most important special case of the single-index model is when the index is the return on the entire set of assets that can be traded on the market. The single-index model then becomes the *market model*. The nature of the market model can be understood by noting that the S+P 500 index is representative of the market but it is not an index of *all* assets. There are stocks of smaller companies that are not included in the index, nor does it include any financial derivatives or bonds. The concept of the market has to be more general than a stock index. For the present time it will be accepted that a market return can be calculated without making precise how this can be done. A detailed dicussion of the issues involved will be given in Chapter 8.

Assume that the index is the market. To denote this special case, the expected return on the index is denoted $\overline{r}_M$, the variance of this return by σ_M^2, and the beta of asset i by β_i^M.

The market model has two features in addition to those of the single-index model. These are derived from the following fact. Let X_i^M denote the proportion of asset i in the market portfolio. Then, by definition, holding the market portfolio must give the market return so

$$\sum_{i=1}^{N} X_i^M r_{i,t} = r_{M,t}. \tag{6.37}$$

Taking the expectation of (6.37) gives

$$\sum_{i=1}^{N} X_i^M \overline{r}_i = \overline{r}_M. \tag{6.38}$$

The first additional feature of the market model is a restriction on the average value of beta. The weighted-average value of beta across the assets, with the weights being the proportion of each asset in the market portfolio, is

$$\overline{\beta}^M = \sum_{i=1}^{N} X_i^M \beta_i^M \tag{6.39}$$

With the market model the beta for asset i is obtained from

$$\beta_i^M = \frac{\sum_{t=1}^{T} \left(r_{i,t} - \overline{r}_i\right)\left(r_{M,t} - \overline{r}_M\right)}{\sum_{t=1}^{T} \left(r_{M,t} - \overline{r}_M\right)^2}. \tag{6.40}$$

Substituting this expression into the definition of weighted-average beta gives

$$\overline{\beta}^M = \sum_{i=1}^{N} X_i^M \frac{\sum_{t=1}^{T} \left(r_{i,t} - \overline{r}_i\right)\left(r_{M,t} - \overline{r}_M\right)}{\sum_{t=1}^{T} \left(r_{M,t} - \overline{r}_M\right)^2}. \tag{6.41}$$

But (6.41) can be written as

$$\overline{\beta}^M = \frac{\sum_{t=1}^{T} \sum_{i=1}^{N} X_i^M \left(r_{i,t} - \overline{r}_i\right)\left(r_{M,t} - \overline{r}_M\right)}{\sum_{t=1}^{T} \left(r_{M,t} - \overline{r}_M\right)^2}. \tag{6.42}$$

However, (6.37) and (6.38) imply that

$$\sum_{i=1}^{N} X_i^M \left(r_{i,t} - \overline{r}_i\right) = \left(r_{M,t} - \overline{r}_M\right). \tag{6.43}$$

Using this results shows that

$$\overline{\beta}^M = \frac{\sum_{t=1}^{T} \left(r_{M,t} - \overline{r}_M\right)\left(r_{M,t} - \overline{r}_M\right)}{\sum_{t=1}^{T} \left(r_{M,t} - \overline{r}_M\right)^2} = 1. \tag{6.44}$$

Therefore, when the market model is used the weighted-average value of β_i^M, with the weights given by the proportions of the assets in the market portfolio, is equal to 1. This is an important result for the interpretation of beta values and for the application of the market model to portfolio analysis.

The fact that the $\overline{\beta}^M = 1$ permits a classification of assets into risk types. An asset that has a value $\beta_i^M < 1$ has less systematic risk than the market portfolio. This means that systematic changes in the return on the asset are smaller, on average, than that change in market return that causes then. Conversely, if $\beta_i^M > 1$ the asset has greater systematic risk than the market. In thiscase, changes in the market return lead, on average, to even greater changes in the return on the asset. These observations provide another reason for viewing the value of beta as a summary measure ofthe riskiness of an asset. They also provide a means to interpret some as being riskier than the market and some as less risky.

The second feature feature of the market model follows from noting that

$$\overline{r}_M = \sum_{i=1}^{N} X_i^M \left[\alpha_i^M + \beta_i^M \overline{r}_M\right]. \tag{6.45}$$

The previous result has shown that $\sum_{i=1}^{N} X_i^M \beta_i^M = \overline{\beta}^M = 1$. This implies

$$\sum_{i=1}^{N} X_i^M \beta_i^M \overline{r}_M = \overline{r}_M \tag{6.46}$$

Substituting (6.46) into (6.45) produces the conclusion that

$$\sum_{i=1}^{N} X_i \alpha_i^M = 0.$$

Hence, the weighted-average value of α_i^M is 0 when the market model is applied.

The next example illustrates these results.

Example 93 *Assume there are only two risky assets available, so these two assets are the entire market. Also assume there are two potential future states of the world and that both states are equally likely. There are* 100 *units of asset*

A and it has an initial price, $p_A(0)$, of 10. There are 200 units of asset B and it has an initial price, $p_B(0)$, of 15. The final prices of the assets in the two states of the world are given in the table.

	State 1	*State 2*
$p_A(1)$	*12*	*11*
$p_B(1)$	*20*	*16*

Given this data, it can be calculated that $\overline{r}_A = 0.15$ and $\overline{r}_B = 0.2$. The proportions of the two assets in the market portfolio can be calculated to be $X_A = 0.25$ and $X_B = 0.75$. Hence the return on the market in state 1 is 0.3 and in state 2 is 0.075. The mean return is $\overline{r}_M = 0.1875$.

Using this data it is possible to calculate β_A^M and β_B^M using the population covariance and variance to evaluate (6.40). Doing this gives

$$\beta_A^M = \frac{\frac{1}{2}(0.2-0.15)(0.3-0.1875)+\frac{1}{2}(0.1-0.15)(0.075-0.1875)}{\frac{1}{2}(0.3-0.1875)^2+\frac{1}{2}(0.075-0.1875)^2} = 0.444,$$

and

$$\beta_B^M = \frac{\frac{1}{2}(0.333-0.2)(0.3-0.1875)+\frac{1}{2}(0.067-0.2)(0.075-0.1875)}{\frac{1}{2}(0.3-0.1875)^2+\frac{1}{2}(0.075-0.1875)^2} = 1.185$$

Using these results the mean value $\overline{\beta}^M$ is

$$\overline{\beta}^M = 0.25 \times 0.444 + 0.75 \times 1.185 = 1.$$

The final step is to compute α_A^M and α_B^M. Using the fact that $\overline{r}_i = \alpha_i^M + \beta_i^M \overline{r}_M$,

$$\alpha_A^M = 0.15 - 0.444 \times 0.1875 = 0.0667,$$

$$\alpha_B^M = 0.2 - 1.185 \times 0.1875 = -0.0222.$$

Hence

$$X_A \alpha_A^M + X_B \alpha_B^M = 0.25 \times 0.0667 + 0.75 \times (-0.0222) = 0.$$

This complete the analysis of the market model. The market model has been identified as a special case of the single-index model. The particular nature of the market model results in two special properties: the weighted-average value of beta is 1 and the weighted-average value of alpha is 0. These properties are not shared by the single-index model in general. These properties permit a classification of assets into those less risky than the market, and those that are more risky. This is useful as a first step in determining the risk of an investment.

6.6 Applying Beta

The beta of an asset plays a very important role in the practical application of investment analysis techniques. The next sections consider it in some detail and develop a practical interpretation of the theory.

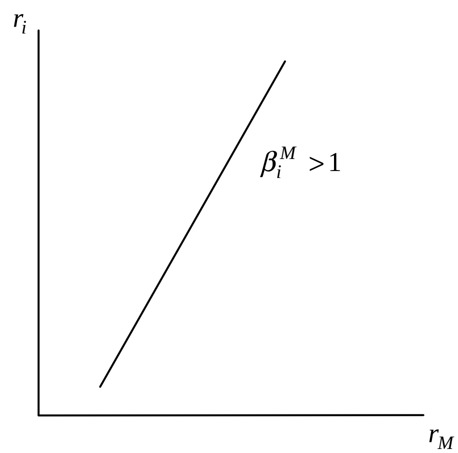


Figure 6.5: Aggressive Asset

6.6.1 Risk

Beta is seen as a measure of the systematic riskiness of an asset. This is clear from (6.22) in which beta can be seen to act as a multiplying factor on the variance of the index. It is also evident that this is not a complete description of risk since the non-systematic risk has also to be taken into account. These statements are also clearly true of a portfolio and the portfolio beta. Even so, this perspective on beta is still helpful.

The observation that beta is related to risk leads to the following interpretations which are given for market model (they can be written equally for the general single-index model):

- If $\beta_i^M > 1$ then asset i is more volatile (or risky) than the market. In this case it is termed "aggressive". An increase (or decrease) in the return on the market is magnified in the increase (or decrease) in the return on the asset. An asset with $\beta_i^M > 1$ is illustrated in Figure 6.5. The figure shows the least squares regression line when the return on an aggressive asset is regressed on the return on the market. It can be seen how changes in r_M are magnified in changes in r_i.

- If $\beta_i^M < 1$ then the asset is less volatile than the market. In this case it is termed "defensive". An increase (or decrease) in the return on the market is diminished in the increase (or decrease) in the return on the asset. An asset with $\beta_i^M < 1$ is illustrated in Figure 6.6. The figure shows the least squares regression line when the return on an defensive asset is regressed on the return on the market. It can be seen how changes in r_M are diminished in changes in r_i.

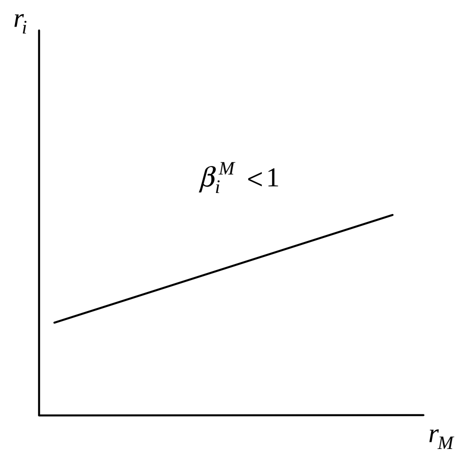


Figure 6.6: Defensive Asset

With these definitions, it is also possible to think in terms of the construction of a "defensive" portfolio of low-beta assets or an "aggressive" portfolio of high-beta assets. Although these are useful descriptions, it should not be forgotten that the total risk must also include the idiosyncratic risk. Only in a well-diversified portfolio can latter be set aside. In a small portfolio with the a poor model of returns the idiosyncratic risk can even dominate the systematic risk.

6.6.2 Adjusting Beta

It has already been noted that beta can be calculated by obtaining historical data on the returns on an asset and on the index. A least squares regression is then conducted of asset returns on index returns. The intercept obtained is α_i and the slope coefficient β_i. There are also several sources for ready-estimated values of beta. Such publications generally provide information on estimates of beta and the non-systematic errors.

The following comments arebest understood by noting that there is now a change in perspective on beta. Up to this point the beta values have been the outcome of a imposing a model on asset returns and estimating a regression line. The practical use of betas in investment analysis often takes the alternative perspective of viewing the beta value as something intrinsic to an asset. That is, the beta is seen as a characteristic of an asset that is (partially) revealed through the data. The role of investment analysis is then to uncover the true value of beta.

This perspective leads to the conclusion that any statistical method for estimating beta will not necessarily obtain the true value. Any data set is but one observation from the set of possible observations, and an estimate is always just the best attempt at measurement given the available data. Any method of computing beta from data then raises some questions about the accuracy of the value obtained and suggests that it may be necessary to adjust the estimated value. Two arguments can be advanced for making an adjustment to the estimated value.

First, any estimate is based on historical data. This can only porduce an accurate value if the value of beta is constant over time. If beta changes, the estimates will be imprecise. Reasons why beta might change concern issues such as the growth and development of the firm, changes in the sturucture of the industry, or maturity of the market. This suggests the argument that the value of an estimated beta should be adjusted to give more emphasis to more recent data.

Second, the estimated value is a random variable. The data is only one realisation of what might have happened, and the value of the estimated beta is dependent on the realisation. If the index used is the market return, then the average value of beta must equal 1. This follows since the average value is the beta of the market and the market has a beta value of 1. The estimated beta is a random variable which is expected (if unbiased) to be equal to the true value. Therefore if the estimated beta deviates from the expected value of 1 there can be two reasons for this. Firstly, the true beta may be different to 1 or, secondly, there is a random error in the estimation. The further the value is from 1 the more likely it is that there is a large random error in the estimation. This suggests that betas that deviate far from 1 may involve large random errors.

This isolates two reasons for considering adjusting estimated betas. Firstly, the value of beta for the stock may change during the course of the data period. Secondly, there is a statistical tendency for large deviations from 1 to be associated with large random errors. The following subsections consider methods of adjustment that can be used to correct estimated values of beta.

6.6.3 Statistical Adjustment

The first two methods of adjustment that are discussed are purely statistical methods. They employ mechanical procedures to make the necessary adjustments to beta.

The analysis of Blume involved estimating beta for a set of stocks for two sample periods, with one period pre-dating the other. The second set of estimates were then regressed on the first set in order to find the average relationship between the betas estimated for the two periods. This process is intended to capture the tendency for mean-reversion in the estimates.

Letting β_{i1} denote the value of beta for stock i in the period 1948-1954 and β_{i2} the value for 1955-1961, the relationship between the two was found to be

$$\beta_{i2} = 0.343 + 0.677\beta_{i1}. \tag{6.47}$$

This result shows clearly the mean-reverting tendency of beta. It also suggests a case for correcting downwards any observed value of beta greater than 1 and adjusting upwards any less than 1.

The correction suggested by Blume is a linear one. It does not put any special emphasis on the sampling error (the extent of deviation from 1) of the observed beta. The Vasilek method is an attempt to do this.

Let $\sigma^2_{\overline{\beta}_1}$ denote the variance of the distribution of the historical estimates of beta over a sample of stock and $\sigma^2_{\beta_i 1}$ be the square of the standard error of the estimate of beta for security i measured in time period 1. Vasilek suggested that an estimate of β_{i2} should be obtained as a weighted average of β_{i1} and $\overline{\beta}_1$, where $\overline{\beta}_1$ is the mean estimate of beta in period 1. The weighting suggested was

$$\beta_{i2} = \frac{\sigma^2_{\overline{\beta}_1}}{\sigma^2_{\overline{\beta}_1} + \sigma^2_{\beta_i 1}} \overline{\beta}_1 + \frac{\sigma^2_{\overline{\beta}_1}}{\sigma^2_{\overline{\beta}_1} + \sigma^2_{\beta_i 1}} \beta_{i1}. \tag{6.48}$$

This weighting procedure adjusts observations with large standard errors further towards the mean than it adjusts observations with small standard errors. It also ensures that the more uncertain is an estimate, the less weight is placed upon it.

6.6.4 Fundamental Beta

The previous section looked at mechanical methods of adjusting beta. In contrast, fundamental betas regard beta as a measure of risk that can be related to the firm-level variables. The basic view is that small, new and indebted firms are more risky.

Particular variables that can be considered are:

- Dividend payout. Often measured by dividends divided by earnings. Since management is more reluctant to cut dividends than to raise them, a high dividend payout is indicative of confidence on the part of management concerning the level of future earnings. Also, dividend payments are less risky than capital gains. Hence, the company that pays more of its earnings as dividends is less risky.

- Asset growth. Often measured by annual change in total assets. Growth is usually though of as positively associated with beta. High-growth firms are thought of as more risky than low-growth firms.

- Leverage. Often measured as senior securities divided by total assets. Leverage tends to increase the volatility of the earnings stream and hence increases risk and beta.

- Liquidity. Often measured as senior securities divided by current liabilities. A firm with high liquidity is thought to be less risky than one with low liquidity and hence liquidity should be negatively related to beta.

- Asset size. Measured by total assets. Large firms are often thought to be less risky than small firms, if for no other reason than that they have better access to the capital markets. Hence they should have lower betas.

- Earning variability. Measured as the standard deviation of the earning/price ratio. The more variable a company's earning stream and the more highly correlated it is with the market, the higher its beta should be.

Given these factors for each firm, the role of the analyst is to subjectively judge how they can be compounded into a value of beta. A standard process would be to start with an estimated beta and then adjust it if it appears to be far out of line on any of these fundamental factors.

6.7 Conclusions

The calculation of the variance of the return for even a medium-sized portfolio can be informationally demanding. The single-index model is a means of reducing the information required. It assumes that a single variable is responsible for generating the returns on all assets. The most important implication of this assumption is that it greatly simplifies the calculation of the variance of the return on a portfolio. Furthermore, it follows that the variance can be decomposed into systematic and non-systematic componentes.

The beta values generated by the single-index model can also be used to categorise assets as aggressive or defensive and provide a simple way of thinking about portfolio construction. Since the betas are estimates, justifications were given for adjusting the estimated value. This lead into a discussion of adjustment methods and fundamental betas.

Exercise 54 *You manage a portfolio of 50 assets and wish to calculate the efficient frontier. If you decide that a sample of 30 observations is required to calculate each variance and covariance, how many data points do you need in total?*

Exercise 55 *One response to the data requirements may be to group stocks into industries and assume that all firms in an industry have the same covariance with all firms from another industry. A variance can then be calculated for each stock and a single covariance. By considering the Ford, General Motors and Dell stock, assess the success of this approach.*

Exercise 56 *Given the following observed returns on an asset and an index, estimate the value of* α *and* β*.*

Period	1	2	3	4	5	6	7	8	9	10
Asset	12	8	5	9	7	15	16	4	3	9
Index	8	7	9	8	12	16	15	7	6	8

Exercise 57 *The following table provides data on the returns on two assets and an index. Assess whether the single-index model is appropriate for these assets.*

Period	1	2	3	4	5	6	7	8	9	10
Asset 1	6	3	6	8	4	4	2	9	4	5
Asset 2	7	8	4	3	6	8	9	4	8	1
Index	2	4	3	9	5	2	8	4	7	1

Exercise 58 *Assume returns are generated by a model where the market is the single factor. The details of the model for three stocks are:*

Stock	Alpha	Beta	$\sigma_{\varepsilon i}$	Portfolio Proportion
A	.1	1.1	4	0.6
B	-.2	0.9	3	0.2
C	.05	0.8	5	0.2

Exercise 59 *The expected return on the market is 12% with a standard deviation of 18%.*

(i) What is the portfolio's expected rate of return?

(ii) What is the standard deviation of the return on the portfolio?

Exercise 60 *Calculate beta for IBM stock using the return on the Standard and Poor 500 over the last 10 years as the index. (Simplify the calculation by ignoring dividends paid on the index).*

Exercise 61 *Assume there are two stocks, A and B, with* $\beta_A = 1.4$ *and* $\beta_B = 0.8$.

(i) If the mean return on the market portfolio is 10% and the risk-free rate of return is 5%, calculate the mean return of the portfolios consisting of:

a. 75% of stock A and 25% of stock B,

b. 50% of stock A and 50% of stock B,

c. 25% of stock A and 75% of stock B.

(ii) If the idiosyncratic variations of the stocks are $\sigma_{\epsilon A} = 4, \sigma_{\epsilon B} = 2$ *and the variance of the market portfolio is* $\sigma_M^2 = 12$, *calculate the variance of the portfolios in (a), (b), (c).*

(iii) What are the mean return and variance of the portfolios in (ii) if they are 50% financed by borrowing?

Exercise 62 *Assume that two assets constitute the entire market. The possible returns in the three future states of the world, which occur with equal probability, and the initial market proportions are given in the table.*

Asset	Proportion	State 1	State 2	State 3
A	0.4	3	2	5
B	0.6	4	4	6

Exercise 63 *(i) Determine the values of* α *and* β *for both assets.*

(ii) Determine the idiosyncratic errors.

(iii) Plot the portfolio frontier.

Exercise 64 *If an investor's risk aversion increases, can the average beta value of their portfolio rise?*

Chapter 7

Factor Models

7.1 Introduction

In a factor model, the return on a security is modelled as being determined by one or more underlying factors. The single-index, or market model of the previous chapter is an example of a single-factor model. In fact, the terminology "factor" and "index" are used interchangeably.

There is no reason to use only a single factor. For instance, firms in the same industry may have returns that rise and fall together due to some correlating factor unique to that industry. If this is the case, the assumption of the single factor model, that the random errors for any two firms are uncorrelated, is not valid.

In general, additional factors may improve the statistical properties of the model and will reduce the unexplained error. Two issues are explored here. First, the returns and variance of a portfolio are derived for models with multiple factors. Second, the set of relevant factors is considered.

7.2 Single-Factor Model

Repeating the definition of the previous chapter, but with a new notation for factors models in general, the returns process for the single-factor model is

$$r_i = a_i + b_i f + e_i \tag{7.1}$$

where f is the single factor.

Repeating the derivations for the market model gives an expected return for asset i

$$\overline{r}_i = a_i + b_i \overline{f}, \tag{7.2}$$

and variance

$$\sigma_i^2 = b_i^2 \sigma_f^2 + \sigma_{e_i}^2, \tag{7.3}$$

where

$$b_i = \frac{\sigma_{if}}{\sigma_f^2}. \tag{7.4}$$

The covariance between two assets i and j is $\sigma_{ij} = b_i b_j \sigma_f^2$.

For a portfolio the return is

$$r_p = a_p + b_p f + e_p, \tag{7.5}$$

and the variance

$$\sigma_p^2 = b_p^2 \sigma_f^2 + \sigma_{e_p}^2, \tag{7.6}$$

where $a_p = \sum_{i=1}^n w_i a_i$, $b_p = \sum_{i=1}^n w_i b_i$ and $e_p = \sum_{i=1}^n w_i e_i$.

7.3 Two Factors

The extension to many factors is now considered, beginning with the case of two factors.

If it is assumed that the returns on asset i are determined by two factors and a random error, the return process becomes

$$r_i = a_i + b_{1i} f_1 + b_{2i} f_2 + e_i, \tag{7.7}$$

where f_1 and f_2 are the values of factors 1 and 2. It is assumed that

$$cov\left(e_i, f_k\right) = 0,\ k = 1, 2,\ \text{all } i, \tag{7.8}$$

and

$$cov\left(e_s, e_j\right) = 0,\ \text{all } i, j. \tag{7.9}$$

With this returns process the expected return on asset i becomes

$$\bar{r}_i = a_i + b_{1i} \overline{f}_1 + b_{2i} \overline{f}_2, \tag{7.10}$$

and the variance of the return

$$\begin{aligned} var\left(r_i\right) &= E\left[\left(r_i - \bar{r}_i\right)^2\right] \\ &= E\left[\left(a_i + b_{1i} f_1 + b_{2i} f_2 + e_i - \bar{r}_i\right)^2\right] \\ &= E\left[\left(b_{1i}\left(f_1 - \overline{f}_1\right) + b_{2i}\left(f_2 - \overline{f}_2\right) + e_i\right)^2\right] \\ &= \sum_{k=1}^{2} b_{ki}^2 E\left[\left(f_k - \overline{f}_k\right)^2\right] + 2 b_{1i} b_{2i} E\left[\left(f_1 - \overline{f}_1\right)\left(f_2 - \overline{f}_2\right)\right] \\ &\quad + E\left[e_i\right]^2 \\ &= b_{1i}^2 \sigma_{f_1}^2 + b_{2i}^2 \sigma_{f_2}^2 + 2 b_{1i} b_{2i} \sigma_{f_1 f_2} + \sigma_{e_i}^2. \end{aligned} \tag{7.11}$$

For two assets, i and j the covariance is

$$\begin{aligned}
cov\left(r_i, r_j\right) &= E\left[\left(r_i - \overline{r}_i\right)\left(r_j - \overline{r}_j\right)\right] \\
&= E\left[\left(a_i + \sum_{k=1}^{2} b_{ki} f_k + e_i - \overline{r}_i\right)\left(a_j + \sum_{k=1}^{2} b_{kj} f_k + e_j - \overline{r}_j\right)\right] \\
&= E\left[\left(\sum_{k=1}^{2} b_{ki}\left(f_k - \overline{f}_k\right) + e_i\right)\left(\sum_{k=1}^{2} b_{kj}\left(f_k - \overline{f}_k\right) + e_j\right)\right] \\
&= \sum_{k=1}^{2} b_{ki} b_{kj} E\left[\left(f_k - \overline{f}_k\right)^2\right] \\
&\quad + \left[b_{1i} b_{2j} + b_{2i} b_{1j}\right] E\left[\left(f_1 - \overline{f}_1\right)\left(f_2 - \overline{f}_2\right)\right] \\
&= b_{1i} b_{1j} \sigma_{f_1}^2 + b_{2i} b_{2j} \sigma_{f_2}^2 + \left[b_{1i} b_{2j} + b_{2i} b_{1j}\right] \sigma_{f_1 f_2}. \qquad (7.12)
\end{aligned}$$

The bs can be calculated by a multiple regression of the return on asset i on the values of the factors. This process guarantees that $cov\left(e_i, f_k\right) = 0$, $k = 1, 2$, all i, and $cov\left(e_s, e_j\right) = 0$, all i, j.

It can also be noted that

$$cov\left(r_i, f_1\right) = b_{1i} \sigma_{f_1}^2 + b_{2i} \sigma_{f_1 f_2}, \qquad (7.13)$$

and

$$cov\left(r_i, f_2\right) = b_{1i} \sigma_{f_1 f_2} + b_{2i} \sigma_{f_2}^2. \qquad (7.14)$$

The values of b_{1i} and b_{2i} can then be solved directly from these equations.

7.4 Uncorrelated factors

An important special case arises when the factors are uncorrelated. If they are then

$$cov\left(f_1, f_2\right) = 0. \qquad (7.15)$$

Employing this assumption gives

$$var\left(r_i\right) = b_{1i}^2 \sigma_{f_1}^2 + b_{2i}^2 \sigma_{f_2}^2 + \sigma_{e_i}^2, \qquad (7.16)$$

and

$$cov\left(r_i, r_j\right) = b_{1i} b_{1j} \sigma_{f_1}^2 + b_{2i} b_{2j} \sigma_{f_2}^2. \qquad (7.17)$$

The values of b_{1i} and b_{2i} follow even more immediately when $\sigma_{f_1 f_2} = 0$. In this case

$$cov\left(r_i, f_1\right) = b_{1i} \sigma_{f_1}^2, \qquad (7.18)$$

and

$$cov\,(r_i, f_2) = b_{2i}\sigma^2_{f_2}, \tag{7.19}$$

so the bs can be found directly. Section 11.6 shows how to construct uncorrelated factors.

7.5 Many Factors

These calculations can be extended directly to any number of factors.

With n factors, the returns process is

$$r_i = a_i + \sum_{k=1}^{n} b_{ki} f_k + e_i, \tag{7.20}$$

where $cov\,(f_k, e_i) = 0$ and $cov\,(e_i, e_j) = 0$.

The expected return becomes

$$\bar{r}_i = a_i + + \sum_{k=1}^{n} b_{ki} \overline{f}_k, \tag{7.21}$$

and the variance is

$$var\,(r_i) = \sum_{k=1}^{n} b^2_{ki} \sigma^2_{f_k} + \sum_{k=1}^{n}\sum_{l=1}^{n} b_{ki} b_{li} \sigma_{f_k f_l} + \sigma^2_{e_i}. \tag{7.22}$$

For two assets, i and j the covariance is

$$cov\,(r_i, r_j) = \sum_{k=1}^{n}\sum_{l=1}^{n} b_{ki} b_{lj} \sigma_{f_k f_l}. \tag{7.23}$$

7.6 DIversification

Show how diversification eliminates the idosyncraitc in the limit.

7.7 Constructing uncorrelated factors

The calculations in Section 11.4 show the simplification that is achieved when the factors are uncorrelated. It is always possible to construct uncorrelated factors.

Consider a model with n factors $f_1, ..., f_n$ which are potentially correlated. The aim is to create factors $\widehat{f}_1, ..., \widehat{f}_n$ which are uncorrelated. To do this, take the first factor, f_1, (it does not matter which this is) and define $\widehat{f}_1 \equiv f_1$. Then conduct the regression

$$f_2 = a + b_1 \widehat{f}_1 + e. \tag{7.24}$$

From this define

$$\widehat{f}_2 = f_2 - \left[a + b_1\widehat{f}_1\right] = e. \tag{7.25}$$

By definition of the least squares estimator, the error, e, must be uncorrelated with f_1. It captures that part of f_2 that is unexplained by f_1.

To obtain $\widehat{f}_i$ then regress

$$f_i = a + \sum_{j=1}^{i-1} b_j\widehat{f}_j + e, \tag{7.26}$$

and define

$$\widehat{f}_i = f_i - a - \sum_{j=1}^{i-1} b_j\widehat{f}_j = e. \tag{7.27}$$

The factors $\widehat{f}_1, ..., \widehat{f}_n$ obtained in this way are uncorrelated as required.

Using these uncorrelated factors, the covariance between two assets i and j is

$$\sigma_{ij} = b_{i1}b_{j1}\sigma^2_{\widehat{f}_1} + ... + b_{in}b_{jn}\sigma^2_{\widehat{f}_n}. \tag{7.28}$$

7.8 Factor models

There are a number of alternative factor models which vary in the motivation for the choice of factors. Two of the most significant are now discussed.

7.8.1 Industry factors

These models begin with the single-index model and add factors that capture industry effects.

If the correlation between securities is caused by a market effect and additional industry effects, then the return generating process becomes

$$r_i = a_i + b_{im}\widehat{f}_m + b_{i1}\widehat{f}_1 + ... + b_{iL}\widehat{f}_L + e_i, \tag{7.29}$$

where $\widehat{f}_m$ is the market index and $\widehat{f}_1, ..., \widehat{f}_L$ are (uncorrelated) factors relating to the L industries in which company i operates.

7.8.2 Fundamental factors

A broader range of factors can be introduced. A way of doing this is based on the efficient market argument that current beliefs about future events are already incorporated in asset prices, so it is only unexpected changes that can affect return. Hence the additional factors should capture these unexpected changes.

An example of an index created on the basis of this reasoning includes as factors:

- *Default risk*: the unexpected difference in return between 20-year government bonds and 20-year corporate bonds. Measured as the return on long-term government bonds minus the return on long-term corporate bonds plus half a per cent.
- *The term structure*: the return on long-term government bonds minus the return on a one-month Treasury bill one month in the future.
- *Unexpected deflation*: the rate of inflation expected at the beginning of month minus the actual rate of inflation realized at the end of the month.
- *Growth*: unexpected change in the growth rate in real final sales.
- *Residual market*: the difference between the excess return on the S&P index and the expected excess return.

	f_1 Default	f_2 Term	f_3 Deflation	f_4 Growth	f_5 Market	R^2
Sector						
Cyclical	-1.53	0.55	2.84	-1.04	1.14	0.77
Growth	-2.08	0.58	3.16	-0.92	1.28	0.84
Stable	-1.40	0.68	2.31	-0.22*	0.74	0.73
Oil	-0.63*	0.31	2.19*	-0.83*	1.14	0.50
Utility	-1.06	0.72	1.54	0.23*	0.62	0.67
Transportation	-2.07	0.58	4.45	-1.13	1.37	0.66
Financial	-2.48	1.00	3.20	-0.56*	0.99	0.72

* Not significant at 5% level.

Exercise 65 *Assume that returns of individual securities are generated by the following two-factor model:*

$$r_{it} = a_i + b_i f_{1t} + c_i f_{2t} + e_{it}.$$

The following three portfolios are observed:

	Expected Return	b_i	c_i
A	*25*	*1.0*	*1.0*
B	*10*	*2.0*	*0.5*
C	*20*	*0.5*	*1.5*

(i) Find the relationship between expected returns and factor sensitivities.

(ii) Suppose you can find a portfolio, D, with expected return = 26, $b_D = 3.0$, $c_D = 1.4$

(iii) Explain how you could construct a profitable arbitrage portfolio from securities A, B and C and portfolio D.

Exercise 66 *Assume that stock returns are generated by a two-factor model*

$$r_{it} = a_i + b_i f_{1t} + c_i f_{2t} + e_{it}.$$

Consider the following portfolio:

Stock	b_i	c_i	e_i
A	*0.2*	*1.1*	*0.6*
B	*0.1*	*1.0*	*0.5*
C	*0.3*	*0.9*	*0.4*

Calculate the variance of an equally-weighted portfolio under the following alternative assumptions:

(i) f_1, f_2 *uncorrelated and* e_i, e_j *uncorrelated* $(i \neq j)$.

(ii) $\rho_{f_1 f_2} = -0.5$ *and* e_i, e_j *uncorrelated* $(i \neq j)$.

Part IV

Equilibrium Theory

Chapter 8

The Capital Asset Pricing Model

There are demands and supplies. There is a balance of forces that gives an equilibrium. When balanced the returns have to be in line. Add some assumptions and generate a clear outcome.

8.1 Introduction

An individual investor managing a small portfolio is correct to treat the expected returns and variances of financial assets as data that are beyond their influence. This does not mean they should not try to understand what explains the pattern of returns that the market presents to them. A deeper understanding can only lead to better investment choices.

The perspective now shifts to consideration of explanations for the observed data. Equilibrium models explain the process of investor choice and market clearing that lies behind the observed pattern of asset returns. That higher expected return means higher risk is an established feature of financial markets. An equilibrium model predicts exactly how much more expected return is required to compensate for additional risk.

What a model does is make assumptions about the trading environment, about the assets that are available, and about the investors who are in the market. From these assumptions it derives trading behaviour. Then it considers the demand and supply of financial assets. An equilibrium for the model is a set of asset prices (or, equivalently, a set of asset returns) such that the demand for all assets is equal to the supply. A model is constructed with the expectation that the form of the equilibrium that emerges will reveal deeper insights into the determination of asset prices in financial markets.

An important benefit of an equilibrium model, and of the Capital Asset Pricing Model (CAPM) in particular, is that it allows the evaluation of portfolio performance. The model generates an equilibrium relationship between expected return and risk. If a portfolio delivers a lower level of expected return than predicted by this relationship for its degree of risk then it is a poor portfolio. The CAPM also carries implications in the area of corporate finance. It can be used as a tool in capital budgeting and project analysis.

The CAPM provides an explanation of asset returns using the concept of a financial market equilibrium. An equilibrium is defined by the supply of assets being equal to the demand for assets. A position of equilibrium is reached by the adjustment of asset prices and, hence, the returns on assets. This adjustment occurs through trading behavior. If the expected return on an asset is viewed as high relative to its risk then demand for the asset will exceed supply. The price of the asset will rise, and the expected return will fall until equilibrium is achieved. The particular assumptions about investors' preferences and information made by a model then determines particular features of the equilibrium. As we will see, the CAPM determines very precise equilibrium relationships between the returns on different assets.

The basic assumption of the CAPM is that all investors follow the process of portfolio selection described in the chapters above. That is, they construct the efficient set and choose the portfolio that makes the value of their mean-variance expected utility as high as possible. Some additional assumptions are then added and the implications are then traced.

It is shown that the CAPM leads to especially strong conclusions concerning the pricing of assets in equilibrium. If the model is correct, these can be very useful in guiding investment and evaluating investment decisions. These issues are pursued further in the chapter on portfolio evaluation.

8.2 Assumptions

The set of assumptions upon which the CAPM is based are now described. The interpretation of each assumption is also discussed. The precise set of assumptions are what distinguish the CAPM from alternative models of financial market equilibrium.

The first set of assumptions describe the properties of the assets that are traded.

All assets are marketable This is the basic idea that all assets can be traded so that all investors can buy or sell anything that is available. For the vast majority of assets this an acceptable assumption. How easily an asset can be traded depends upon the extent to which an organized market exists. There are some assets cannot be easily traded. An example is human capital. It can be rented as a labor service but cannot be transferred from one party to another.

All assets are infinitely divisible The consequence of this assumption is that it is possible to hold any portfolio no matter what are the portfolio proportions. In practice assets are sold in discrete units. It is possible to move close to a position of infinite divisibility by assuming the existence of mutual funds that allow investor to purchase fractions of assets. For instance, a treasury bill may have a denomination of $100,000 but a small fraction of one can be bought if it is shared between a large number of investors.

The second set of assumptions characterize the environment in which trade takes place.

No transaction costs Transaction costs are the costs of trading. In practice, brokers charge commission for trade and there is a spread between the buying and selling prices. Margin must also be deposited as security for several forms of deal. The model assumes a financial market in which such costs are absent. The role of the assumption is to allow portfolios to be adjusted costlessly to continually ensure optimality.

Short sales are allowed The effect of introducing short sales has already been shown in the construction of the efficient frontier. Short sales also play an important role in arbitrage. In brief, arbitrage involves an over-priced asset being sold short with the proceeds of the short sale used to buy an under-priced (or efficiently priced) asset or portfolio with the same risk. The risk on the two assets cancels, so a positive return is earned with no risk and no net investment. The role of arbitrage is to ensure that all profitable opportunities are exhausted – which they must be in equilibrium. Short sale are permitted in actual financial markets, so the assumption captures a realistic trade. Where the CAPM diverges from practice is that it is assumed there are no charges for short selling. In practice margin must be deposited with the broker which is costly to the investor since it earns less than the market return.

No taxes Taxes affect the returns on assets and tax rules can alter the benefit of capital gains relative to dividends and coupons. The assumption that there are no taxes removes the distortions from the market.

The next pair of assumptions imply that the financial markets are competitive with no asymmetries of information..

Lending and borrowing can be undertaken at the risk-less rate Investors face a single rate of interest. This is the assumption of a perfect capital market. There are no asymmetries of information that prevent lending and borrowing at a fair rate of interest. The consequence of this assumption is that the efficient frontier is a straight line.

No individual can affect an asset price This is idea of a competitive market where each trader is too small to affect price. It takes away any market power and rules out attempts to distort the market. All investors can plan their trades on the basis of the prices they observe and do not need to be concerned with trades affecting prices.

The next set of assumptions describe the objectives and information of the investors in the model.

All investors have mean-variance preferences This is the assumption that we have already used in the analysis of the Markovitz model. The assumption allows us to set the model in mean variance space and analyze the choice of an optimal portfolio through combining preferences and the efficient frontier.

All investors have a one period horizon This is also one of the assumptions that was imposed in the analysis of the Markovitz model. The assumption simplifies the analysis and representation of the investment decision. An investor is concerned only with the purchase price of an asset and the selling price at the end of the single holding period.

The final assumption of the CAPM is critical for the equilibrium that is obtained. The assumption imposes one dimension of similarity among the individual investors.

All investors hold the same expectations This assumption ensures that all investors have the same information and, on the basis of this information, reach the same conclusions about the distribution of returns for each asset. Note carefully that the assumption does not make all the investors identical. The CAPM allows investors to differ in risk aversion. Some may be very risk averse some may be less risk averse. What the assumption does is ensure that all the investors are equally informed but does not restrict preferences.

Example 94 *Give example of same information and different preferences. Here idea that one investor believes an asset will rise and another that it will fall cannot be compatible with the assumed information structure.*

This set of assumptions combines the Markovitz model of portfolio choice developed in earlier chapters with the assumption that investors have the same information and reach the same assessment of the expected return and variance of return for every asset. It is the information assumption that permits the aggregation of individual choices into a market equilibrium with specific properties.

8.3 Equilibrium Implications

The important properties of equilibrium are now determined by tracing through the implications of the CAPM assumptions. The properties that emerge are what has made the CAPM such a significant model for the theory of financial markets.

In general an equilibrium for a model of investment is a set of asset prices that ensure the demand for every asset is equal to the supply of that asset.

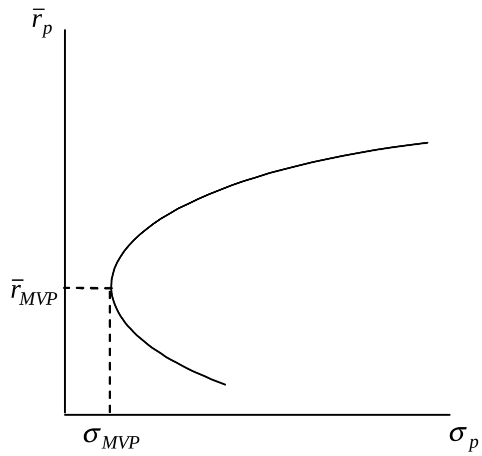


Figure 8.1: Portfolio frontier

The CAPM is a single period model so by a set of prices is meant a set of initial prices at which assets are bought and a set of final prices at the end of the single holding period that capture the final asset values. Since the prices determine the returns, this means also means a set of returns for the assets. It is possible to analyze the CAPM directly in this way. But the CAPM has some very special features that make it more informative to take a "sideways" approach to equilibrium. This uses the assumptions of the model to build a picture of equilibrium.

8.3.1 Separation

The investors all have the same information and expectations. As the investors conform to the Markovitz model of portfolio choice they use this information to construct the portfolio frontier. Having the same expectations it follows that the investors perform the same financial calculations. Hence, all investors construct the same portfolio frontier for risky assets and assess there to be the same trade-off between expected return and risk.

The general form of portfolio frontier for the risky assets constructed by all investors is shown in Figure 8.1. The key features of this frontier are the identification of the minimum variance portfolio and that when there are more than two risky assets the frontier is the outer envelope of the portfolio set. That is, there are inefficient portfolios that lie inside the frontier.

By assumption all the investors face the same risk-free rate of return. Combining this with the portfolio frontier implies that the investors must all construct the same efficient frontier. In particular, the tangency portfolio must be the same for all investors. This situation is illustrated in Figure 8.2 where r_f is the risk-free rate of return, point M is the tangency portfolio, and MVP denotes the minimum variance portfolio.

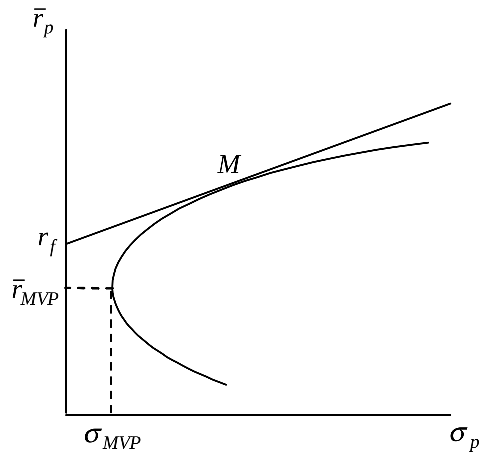


Figure 8.2: Efficient frontier and market portfolio

All of the investors choose their optimal portfolio from this efficient frontier. The analysis of the Markovitz model showed that an investor faced with a linear frontier will choose a portfolio that combines the tangency portfolio of risky assets, M, and the risk-free asset. However, the investors can differ in preferences. As a consequence, the proportions in which the tangency portfolio and the risk-free asset are combined will differ according to the degree of risk aversion of each investor. The more risk averse investors will lend at the risk-free rate. Investors who are less risk averse may borrow at the risk-free rate to raise the investment in the tangency portfolio.

Example 95 *Give example of this market portfolio for a simple economy.*

Since all consumers are purchasing portfolio M, this must be the *market portfolio* of risky assets. The term market portfolio refers to a portfolio that combines the risky assets in the same proportions as they are found in the market as a whole. The tangency portfolio has to be the market portfolio since no investor purchases any other combination of risky assets. The sum over all investors of the number of units of the individual assets in these tangency portfolios must sum to the total number that is available. This is the fundamental idea of supply being equal to demand for the individual assets. This is what was meant by a sideways approach to equilibrium. The analysis does not demonstrate the supply equals demand condition explicitly but derives how it must be structured from the observation that only the tangency portfolio is purchased.

What is behind this argument is that the prices of the assets, and the implied rates of return, adjust to make the individual assets more or less attractive until the point is reached at which the demands and supplies are equal. This is also true of the risk-free asset. The model permits there to be borrowing and lending. The risk-free rate of return is adjusted until the total of borrowing is the same

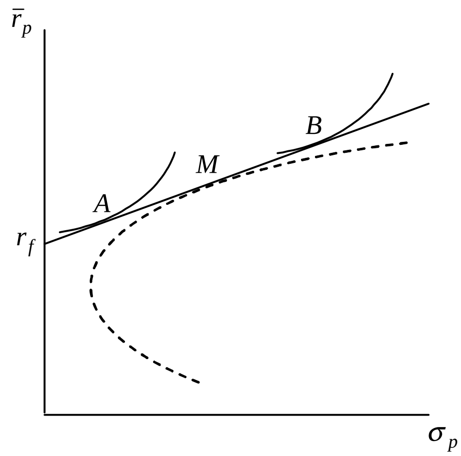


Figure 8.3: Separation principle

as the total of lending. None of this is explicit in this presentation of the model but is in the background behind the arguments.

These conclusions give rise to the *separation principle.* This principle states that an investor only needs to purchase two different assets. The first asset is the risk-free asset. The second asset is the market portfolio. The market portfolio could be marketed as an indirect investment into a mutual fund that holds the market portfolio. What differs between the individual investors is not the range of assets that are held but the proportions of these assets in the optimal choice of portfolio. The more risk averse is an investor the higher will be the proportion of the risk free asset in the portfolio. Less risk averse investors will hold a larger proportion of the market portfolio. Those with a low enough level of risk aversion will go short in the risk free asset to invest in the market portfolio.

The separation principle is illustrated in Figure 8.3. The investor who chooses portfolio A is more risk-averse than the investor that chooses portfolio B. Portfolio A combines a positive holding of the risk-free asset ($X_f > 0$) with a positive holding of the market portfolio ($X_M > 0$). The investment in the risk-free asset involves lending to the issuer. In contrast, portfolio B involves going short in the risk-free asset ($X_f < 0$) and correspondingly long in the market portfolio ($X_M > 1$). Going short in the risk-free implies borrowing to invest, which is the process of buying on the margin. The content of the separation principle is that the needs of both investors are met using just the risk-free asset and the market portfolio. Even though the investors differ in attitudes to risk their investment needs are served by just two assets.

There is one further implication of the separation principle. If all investors are purchasing the same tangency portfolio of risky assets it follows that there can be no short selling in equilibrium. This is because if any investor were short-selling a risky asset it would mean that all would be short-selling the asset. This cannot be an equilibrium since short selling is negative demand, so the aggregate demand for the asset would be negative. A negative demand cannot be equal to a positive supply, so this cannot be an equilibrium for the financial markets.

The analysis of the CAPM can be simplified by assuming that the market portfolio is well-diversified. This assumption can then be used to argue that non-systematic risk is diversified away by all investors since they hold the well-diversified market portfolio. A justification for this additional assumption can be obtained by drawing an analogy with the actual financial market in which literally thousands of securities can be traded. This assumption provides some helpful simplification but is not necessary for deriving any of the key results. The equilibrium of the CAPM that has just been described, and the further properties that will be discussed next, are true even if the financial market consists of just two risky assets. With just two assets the diversification argument cannot be applied so non-systematic risk remains. Even so, the equilbrium properties of the CAPM are unaffected.

8.3.2 Capital Market Line

The assumptions of the CAPM imply that all investors choose from the same efficient frontier. If the CAPM model is to have any predictive content it is necessary to be able to determine the structure of the efficient frontier. The purpose of this section is to show that the efficient frontier in the CAPM does have a special structure. This structure is a consequence of the market portfolio being the tangency portfolio.

The structure of the efficient frontier can be constructed as follows. First, it is always the case that the risk-free asset is on the efficient frontier. Second, the equilibrium of the CAPM implies that market portfolio is always on the efficient frontier. These observation impy that the efficient frontier, when plotted with standard deviation on the horizontal axis and expected return on the vertical axis, connects the points $(0, r_f)$ (the risk-free) and $(\sigma_M, \overline{r}_M)$ (the market portfolio). The first point determines the intercept of the efficient frontier with the vertical axis. The two points together determine that the gradient, g,of the efficient frontier is

$$g = \frac{\bar{r}_M - r_f}{\sigma_M - 0} = \frac{\bar{r}_M - r_f}{\sigma_M}. \tag{8.1}$$

Using these observations it follows that expected return and the standard deviation of any portfolio, p, that lies on the efficient frontier must satisfy the equation

$$\bar{r}_p = r_f + \left[\frac{\bar{r}_M - r_f}{\sigma_M}\right] \sigma_p. \tag{8.2}$$

This special form efficient frontier is usually call the *capital market line* and implies that there is a linear relationship between risk and expected return for all efficient portfolios that might be chosen by an investor in equilibrium. .

The capital market line is illustrated in Figure 8.4. The line passes through the locations of the market portfolio and the risk-free asset. In equilibrium all

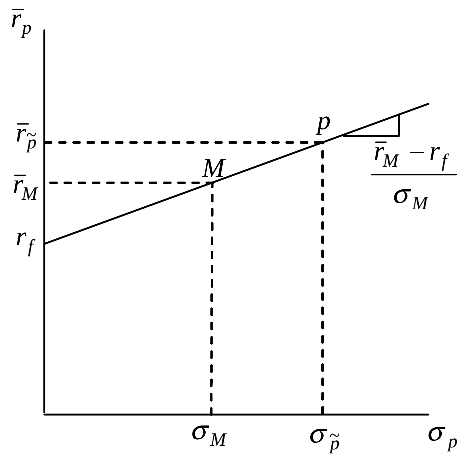


Figure 8.4: Capital Market Line

efficient portfolios have risk and expected return combinations that lie on this line. The risk and expected return combination of one efficient portfolio, $\tilde{p}$, is shown in the figure.

The interpretation of (8.2) is that the intercept r_f is the reward for "time". This is the return earned by an investor who holds the risk-free asset. This involves no risk, but the investor still requires compensation for the postponement of consumption. Holding the risk-free asset delays consumption for one period and the compensation received is the risk-free rate of return. The gradient of the capital market line, $\frac{\bar{r}_M - r_f}{\sigma_M}$, is interpreted as the reward for "risk" and is often called the *market price of risk*. This can be explained by noting that a risk-averse investor requires compensation beyond that given by the risk-free rate in order to hold a risky portfolio. In the equilibrium of the CAPM the capital market line shows that each unit of standard deviation is rewarded by an extra $\frac{\bar{r}_M - r_f}{\sigma_M}$ units of expected return. The term $\frac{\bar{r}_M - r_f}{\sigma_M}$ is called the *Sharpe ratio* and is used the process of portfolio evaluation described in Chapter 17.

Example 96 *Assume $r_f, \bar{r}_M$ and σ_M. The construct capital market line. Then take an asset σ_i and find the implied $\bar{r}_i$.*

Before proceeding it must be stressed that the returns on risky portfolios are, by definition, random. It is easy to fall into the trap of thinking that the return on a portfolio should always be on the capital market line given its level of risk. This is wrong since it is the *expected* return of the portfolio that locates it on the capital market line and not the *realized* return in any particular period. In any particular period the realized portfolio return of a risky portfolio may be above or below the value predicted by the capital market line given its risk. Only in expected terms is the return located upon the capital market line. This is just to re-stating that randomness distinguishes r_p from $\bar{r}_p$. If the realized returns of a set of portfolios in a particular time period are plotted then there should be a random distribution of points above and below the capital market line.

Example 97 *Need an example that returns to the state of the world model to show precisely what is meant here.*

These observations provide a guide to trading behaviour. The question that can be asked is whether the expected return of a portfolio higher or lower than that predicted by the capital market line. If it is higher, then the advice from applying the model is to buy the portfolio. If it is below, then the advice from the model is to sell. If the model is exactly true and provides a correct description of the financial market it should not be possible to discover a portfolio that is located above the capital market line. However, if the model is almost true (for instance, financial markets have some transaction costs or other impediment to costless trade), it may take time for the market to reach a new equilibrium after some fundamental change has occured. During a period of temporary disequilbrium some portfolios can be above the line.

Example 98 *Now talk of trading strategies and evaluation of portfolio performance.*

8.3.3 Security Market Line

The capital market line has shown how the expected return on an efficient portfolio is related to risk. For individual assets the single-index model revealed the benefits from summarizing risk through the use of the beta coefficient. The CAPM also permits the returns on individual assets to be expressed in terms of a beta coefficient. This result can be obtained by modifying the way in which the capital market line presents the equilibrium implications. This will give a very simple expression relating returns and risks for individual assets. Two derivations of this result are given. The first is a simple intuitive derivation. The second derives the result formally by constructing the efficient frontier.

The basic insight is obtained by graphing the covariance of the return on an asset with the return on the market (σ_{iM}) against the expected return of the asset ($\bar{r}_i$). This is the space in which a beta value can be constructed. Two points on the graph can be obtained by considering the risk-free asset and the market portfolio.

The covariance of the risk-free asset with the market is zero and its return is r_f. This gives the point $(0, r_f)$ which is the intercept of the graph with the vertical axis. The covariance of the market with the market is σ_M^2, and the expected return is $\bar{r}_M$. This gives the second point on the graph $\left(\sigma_M^2, \bar{r}_M\right)$. Combining the market portfolio and the risk-free asset permits movement along a line through these two points. This is the *Security Market Line*.

Using the two identified points, the equation of the security market line is

$$\bar{r}_i = r_f + \left[\frac{\bar{r}_M - r_f}{\sigma_M^2}\right]\sigma_{iM}, \tag{8.3}$$

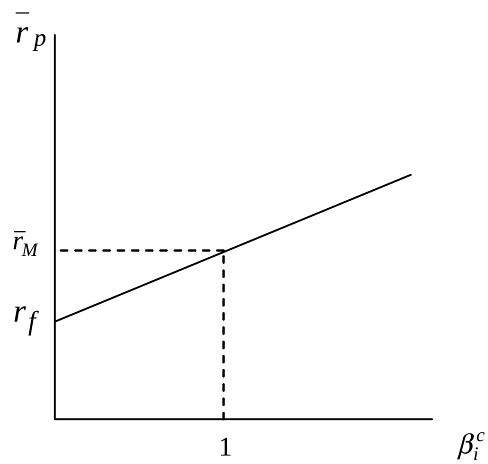


Figure 8.5: Security market line

or, defining $\beta_i^C = \frac{\sigma_{iM}}{\sigma_M^2}$, (here the superscript C denotes the CAPM model)

$$\bar{r}_i = r_f + [\bar{r}_M - r_f] \beta_i^C. \tag{8.4}$$

Hence there is a linear trade-off between risk measured by β_i^C and return $\bar{r}_i$.The security market line is illustrated in Figure 8.5.

In equilibrium all assets must offer return and risk combinations that lie on the security market line. If there was an asset located above this line, all investors would buy it. Equally, if there was an asset that lay below the line, no investor would hold it. Trade in the assets must ensure that in equilibrium they will lie on the line. It is important to think carefully about these arguments. The model is one of equilibrium and the Security Market Line has been constructed as a property of equilibrium. There is an inconsistency with then using it to determine what to do in a disequilibrium position..But this conflict can be resolved if it is thought that there can be short periods of temporary disequilibrium that allow returns on different assets to become slightly out of line. And it is picking the assets whose returns are temporarily high that constitutes a successful investment strategy.

Example 99 *Give some examples. Best to use S+P as the market and do some asset calculation.*

Example 100 *of buying and selling*

The formal derivation of the security market line from the efficient frontier is now given. It is first necessary to establish a preliminary result. Let the portfolio proportions in the market portfolio be $\left(X_1^M, \ldots, X_N^M\right)$. For any asset

i it must then be true that

$$\sigma_{iM} = \sum_{j=1}^{N} X_j^M \sigma_{ji}. \tag{8.5}$$

To demonstrate this result observe that

$$\begin{aligned} \sigma_{iM} &= E\left[(r_i - \bar{r}_i)(r_M - \bar{r}_M)\right] \\ &= E\left[(r_i - \bar{r}_i)\left(\sum_{j=1}^{N} X_j^M r_j - \sum_{j=1}^{N} X_j^M \bar{r}_j\right)\right] \\ &= E\left[(r_i - \bar{r}_i)\sum_{j=1}^{N} X_j^M (r_j - \bar{r}_j)\right]. \end{aligned} \tag{8.6}$$

The expectations operator allows summation to be taken outside so

$$\begin{aligned} E\left[(r_i - \bar{r}_i)\sum_{j=1}^{N} X_j^M (r_j - \bar{r}_j)\right] &= \sum_{j=1}^{N} X_j^M E\left[(r_i - \bar{r}_i)(r_j - \bar{r}_j)\right] \\ &= \sum_{j=1}^{N} X_j^M \sigma_{ji}, \end{aligned} \tag{8.7}$$

which is the result that had to be demonstrated.

The efficient frontier is obtained by choosing the portfolio that maximizes the gradient of the line joining the risk-free asset to the location of the portolio on the frontier for risky assets. This determines the tangency portfolio. From (4.22) the gradient of the line joining the risk-free to portfolio p is

$$g = \frac{\bar{r}_p - r_f}{\sigma_p}. \tag{8.8}$$

Written explicitly in terms of the weights describing portfolio p the gradient is

$$g = \frac{\sum_{j=1}^{N} X_j \bar{r}_j - r_f}{\left(\sum_{j=1}^{N}\sum_{k=1}^{N} X_j X_k \sigma_{jk}\right)^{1/2}}. \tag{8.9}$$

The gradient has to be maximized subject to the the constraint that the portfolio weights sum to 1

$$\sum_{j=1}^{N} X_j = 1. \tag{8.10}$$

The constraint can be substituted into the objective to give the unconstrained optimization

$$\max_{\{X_1,\ldots,X_N\}} \frac{\sum_{j=1}^{N} X_j \left[\bar{r}_j - r_f\right]}{\left(\sum_{j=1}^{N}\sum_{k=1}^{N} X_j X_k \sigma_{jk}\right)^{1/2}}, \tag{8.11}$$

where (8.11) follows from (8.9) and (8.10) because $r_f = \sum_{j=1}^{N} X_j r_f$ only when $\sum_{j=1}^{N} X_j = 1$. The necessary condition for the choice of X_i is

$$[\bar{r}_i - r_f]\left(\sum_{j=1}^{N}\sum_{k=1}^{N} X_j X_k \sigma_{jk}\right)^{-1/2} + \sum_{j=1}^{N} X_j [\bar{r}_j - r_f]\left(-\frac{1}{2}\right)\left(\sum_{j=1}^{N}\sum_{k=1}^{N} X_j X_k \sigma_{jk}\right)^{-3/2} 2\left(\sum_{j=1}^{N} X_j \sigma_{ji}\right) = 0.$$

Now observe that the tangent portfolio is the market portfolio in the CAPM. This implies that

$$\sigma_M^2 = \sum_{j=1}^{N}\sum_{k=1}^{N} X_j^M X_k^M \sigma_{jk}, \tag{8.12}$$

and that

$$\sum_{j=1}^{N} X_j^M [\bar{r}_j - r_f] = \bar{r}_M - r_f. \tag{8.13}$$

Substituting these result and (8.5) into (**??**) allows the necessary condition to be written as

$$[\bar{r}_i - r_f] - [\bar{r}_M - r_f]\frac{\sigma_{iM}}{\sigma_M^2} = 0. \tag{8.14}$$

Using $\beta_i^C = \frac{\sigma_{iM}}{\sigma_M^2}$ an rearranging gives

$$\bar{r}_i = r_f + \beta_i^C [\bar{r}_M - r_f], \tag{8.15}$$

which is the Security Market Line.

8.4 CAPM and Single-Index

The CAPM and the single-index model both generate a parameter β which determines the return on the asset. Consequently, it is important to make clear the difference in meaning and interpretation of β_i and β_i^C.

The basic difference is that β_i is derived from an assumption about the determination of returns. In particular, it is derived from a statistical model of the return process. The index on which returns are based is chosen, not specified by any underlying analysis. In contrast, β_i^C is derived from an equilibrium theory. It emerges from the assumptions of that theory rather than being imposed upon it. The assumptions also generate a precisely defined value for β_i^C.

Also, in the single-index model, the index is usually assumed to be an index representing the market as a whole, but in principal could be any index. In the CAPM, M is always the market portfolio. Note now that there are even closer links between the market model and the CAPM. In both cases the covariance is with the return on the market portfolio. So in principal could be the same, but the market model can be applied even if not all the assumptions of CAPM are true. Finally, the CAPM provides a sufficient set off assumptions for the single-index model to be the true representation of the return-generating process rather than just an approximation. Under its assumptions, returns are generated by a linear relationship.

The relationship between the two models can be understood further by considering the estimation process involved. If the single index is fitted to data using the return on the market portfolio as the index the estimating equation is

$$r_i = \alpha_i^M + \beta_i^M r_M + \varepsilon_i^M. \tag{8.16}$$

The equation that is estimated for the CAPM is defined in terms of the return in excess of the risk-free rate. So the CAPM equation is

$$r_i - r = \alpha_i^C + \beta_i^C [r_M - r] + \varepsilon_i^C. \tag{8.17}$$

Now notice that the equation for CAPM can be transformed to give

$$r_i = \alpha_i^C + \left[1 - \beta_i^C\right] r + \beta_i^C r_M + \varepsilon_i^C. \tag{8.18}$$

It can now be seen from comparison of (8.16) and (8.18) that the two regression will produce the same values of beta, so $\beta_i^M = \beta_i^C$. In addition, it will also be the case that

$$\alpha_i^M = \alpha_i^C + \left[1 - \beta_i^C\right] r. \tag{8.19}$$

This latter equality can be refined further. Take the expectation of both sides of (8.17) to obatin

$$E[r_i - r] = \alpha_i^C + \beta_i^C E[r_M - r]. \tag{8.20}$$

But the security market line (8.15) asserts that

$$E[r_i - r] = \beta_i^C E[r_M - r]. \tag{8.21}$$

Consequently,

$$\alpha_i^C = 0.$$

This leads to the conclusion that if the regression for the single-index model is run with the return on the market as the index then

$$\alpha_i^M = \left[1 - \beta_i^M\right] r. \tag{8.22}$$

The values of α_i^M are determined by the asset's beta and the risk-free rate.

The fact that $\alpha_i^C = 0$ if the CAPM applies forms the basis for tests of the model. This is undertaken by running a regression of (8.17) and conducting a hypothesis test of $\alpha_i^C = 0$. The empirical testing is described in detail in Chapter 10.

Example 101 *But of what? Show that the two will be equal. I.e. used the two methods of calculation to show the same. That is use the basic relationship that*

$$\beta_i = \frac{cov\left(r_i, r_m\right)}{var\left(r_m\right)} = \frac{cov\left(r_i - r_f, r_m - r_f\right)}{var\left(r_m - r_f\right)}$$

Example 102 *Use the general relationship to demonstrate that* $cov\left(r_i, r_m\right) = cov\left(r_i - r_f, r_m - r_f\right)$ *and* $var\left(r_m\right) = var\left(r_m - r_f\right)$.

8.5 Pricing and Discounting

Have so far said nothing that has been specific about prices. Now remedy this. Once it has been done it provides an approach to discounting in general that can be applied not just to assets but to a whole range of risky projects.

8.5.1 Prices

The CAPM also has implications for asset prices. Since the returns of assets are related by the Security Market Line in equilibrium, the prices must also be related.

Recall the role of prices. There are initial purchase prices of the assets at the start of the single holding period. There are final values of the assets at the end of the holding period. The initial prices are known when the assets are bought. The final prices are uncertain for the risky assets. If they where certain then the returns would be known and the asset would not be risky. It has been assumed that investors form expectations about returns. This is equivalent to saying that they form expectations about the asset prices at the end of the holding period. Since it is the expected returns that determine the pattern of asset holdings it is not surprising that the model can be written in terms of expected prices.

To derive the relationship for asset prices, note that the return on an asset can be written as

$$r_i = \frac{q_i - p_i}{p_i}, \tag{8.23}$$

where p_i is the purchase price and q_i the (random) sale price. If dividends are paid, they can be incorporated within q_i. From the security market line

$$\overline{r}_i = r_f + \beta_i^C \left[\overline{r}_M - r_f\right]. \tag{8.24}$$

So

$$\frac{\overline{p}_i\left(1\right) - p_i\left(0\right)}{p_i\left(0\right)} = r_f + \beta_i^C \left[\overline{r}_M - r_f\right], \tag{8.25}$$

or

$$p_i\left(0\right) = \frac{\overline{p}_i\left(1\right)}{1 + r_f + \beta_i^C \left[\overline{r}_M - r_f\right]}. \tag{8.26}$$

This should be the equilibrium market price of the asset.

Note role here: work out expected price and dividend in period 1 and discount back to period 0. The role of β_i^C is to adjust the risk free rate of return to give the correct rate of discounting for the degree of risk of the asset.

8.5.2 Discounting

This illustrates a general principle for discounting to find the present value of a project. Note that $\overline{p}_i(1)$ can just be seen as the expected value of a future random payoff from any kind of investment project. Then $p_i(0)$, the value today, is just the discounted value of the that set of payments. The discounting includes the return on risk-free to represent the time element and the beta term to reflect correction for risk. Notice that the higher is beta the greater is the discounting. So more risky projects (more risky in terms of beta with market) are discounted more heavily.

To see this as a general process observe that the problem at the heart of valuation is to take a sequence of random cash flows $\left\{\widetilde{C}_t\right\}$, $t = 0, ..., T$, and to construct a present value at time 0. If preferences are risk neutral, the present value is found easily by discounted the expected cash flow at t and discounting at the risk-free rate. This would give

$$PV_0 = C_0 + \frac{E\left(\widetilde{C}_1\right)}{1 + r_f} + \frac{E\left(\widetilde{C}_2\right)}{\left[1 + r_f\right]^2} + \ldots, \tag{8.27}$$

where C_0 is taken as known at time 0. The difficulties begin when there is risk aversion. Several methods are now considered for achieving the valuation with risk aversion.

Discount at a rate capturing the risk in the cash flow. The present value then becomes

$$PV_0 = C_0 + \frac{E\left(\widetilde{C}_1\right)}{1 + r_c} + \frac{E\left(\widetilde{C}_2\right)}{\left[1 + r_c\right]^2} + \ldots, \tag{8.28}$$

with $r_c = r_f + r_p$. Here r_p can be interpreted as the risk premium that the the risky cash flow must pay in excess of the risk-free rate. The difficulty in using this approach is the determination of r_p. It should reflect the premium applied to other assets with similar risk.

Use the Certainty Equivalent. For each random cash flow there is a certainty equivalent that satisfies

$$U\left(C_t^e\right) = EU\left(\widetilde{C}_t\right), \tag{8.29}$$

so that the utility of the certainty equivalent is equal to the expected utility of the random cash flow. The present value then becomes

$$PV_0 = C_0 + \frac{C_1^e}{1 + r_f} + \frac{C_2^e}{\left[1 + r_f\right]^2} + \ldots. \tag{8.30}$$

This method is limited by the need to employ the utility function to determine the certainty equivalent.

Each of these methods will work but has its own drawbacks. A further method is now proposed and then explored in detail. Apply CAPM. The risk premium r_p can be determined very easily if the CAPM model is appropriate. If CAPM applies then the security market line gives the relationship

$$r_c = r_f + \beta_c^C [r_M - r_f]. \tag{8.31}$$

The drawback with using CAPM is that it relies on restrictive assumptions.

Example 103 *Add a simple example of how this can be used. Three states, market return, covariances, payment on project.*

Example 104 *Next example give market variance, project covariance, expected value. Find beta and value project.*

8.6 Market Portfolio

The CAPM model relies on the use of a market portfolio in order to be operative. This market portfolio is meant to the entire set of risky assets that are available. It is not clear how this is obtained.@@@AN EXPANDED discussion is necessary.

The major difficulty is the breadth of the market portfolio. It is meant to include all risky assets not just financial securities. For example, it includes real assets such as art and property and other assets such as human capital. This is obviously not easy to define.

There are three situations in which this problem of defining the market portfolio arises. The first is in the calculation of the beta values for assets. Recall that these are obtained by covariance of the return on an asset with the market divided by the variance of the return on the market. If the market portfolio is incorrectly defined both of these values will also be wrong and the estimated beta will not be correct.

The next problem is the construction of the capital market line and the security market line. If an incorrect market portfolio is chosen and the beta values estimated on the basis of this are wrong then the two lines will not provide the correct predictions on returns.

The final problem is that the problem of the market portfolio makes it difficult to test whether the CAPM model is correct or not. If the prediction of the security market line is used as a test of the model then a rejection can show that either the model does not apply or the wrong market portfolio is used. More is said about this in Chapter 10.

8.7 Extension of CAPM

ADD brief discussion of extensions.

8.8 Conclusions

The CAPM moves us from fact (the acceptance of returns and variances as data and the analysis of choice) to modelling of where this data comes from. The CAPM determines the returns in equilibrium by assuming that they are determined by adjustment of returns to equate the demand and supply of assets.

CAPM gives very clear conclusions. It explains the returns on assets through the relationship with the market portfolio. It also gives a guide to investment behavior through the combination of the market portfolio and the risk free asset. The model also formalize why betas are of interest in investment analysis. But all of these properties must be confronted with evidence since the assumptions are equally strong.

Exercise 67 *Assume there are two stocks, A and B, with* $\beta_A = 1.4$ *and* $\beta_B = 0.8$. *Assume also that the CAPM model applies.*

(i) If the mean return on the market portfolio is 10% and the risk-free rate of return is 5%, calculate the mean return of the portfolios consisting of:

a. 75% of stock A and 25% of stock B,

b. 50% of stock A and 50% of stock B,

c. 25% of stock A and 75% of stock B.

(ii) If the idiosyncratic variations of the stocks are $\sigma_{\epsilon A} = 4, \sigma_{\epsilon B} = 2$ *and the variance of the market portfolio is* $\sigma_M^2 = 12$, *calculate the variance of the portfolios in (a), (b), (c).*

(iii) What are the mean return and variance of the portfolios if they are 50% financed by borrowing?

Exercise 68 *Assume there are just two risky securities in the market portfolio. Security A, which constitutes 40% of this portfolio, has an expected return of 10% and a standard deviation of 20%. Security B has an expected return of 15% and a standard deviation of 28%. If the correlation between the assets is 0.3 and the risk free rate 5%, calculate the capital market line.*

Exercise 69 *The market portfolio is composed of four securities. Given the following data, calculate the market portfolio's standard deviation.*

Security	*Covariance with market*	*Proportion*
A	*242*	*0.2*
B	*360*	*0.3*
C	*155*	*0.2*
D	*210*	*0.3*

Exercise 70 *Given the following data, calculate the security market line and the betas of the two securities.*

	Expected return	*Correlation with market portfolio*	*Standard deviation*
Security 1	*15.5*	*0.9*	*2*
Security 2	*9.2*	*0.8*	*9*
Market portfolio	*12*	*1*	*12*
Risk free asset	*5*	*0*	*0*

Exercise 71 *Consider an economy with just two assets. The details of these are given below.*

	Number of Shares	Price	Expected Return	Standard Deviation
A	100	1.5	15	15
B	150	2	12	9

The correlation coefficient between the returns on the two assets if 1/3 and there is also a risk free asset. Assume the CAPM model is satisfied.

(i) What is the expected rate of return on the market portfolio?

(ii) What is the standard deviation of the market portfolio?

(iii) What is the beta of stock A?

(iv) What is the risk free rate of return?

(vi) Construct the capital market line and the security market line.

Exercise 72 *Consider an economy with three risky assets. The details of these are given below.*

	No. of Shares	Price	Expected Return	Standard Deviation
A	100	4	8	10
B	300	6	12	14
C	100	5	10	12

The correlation coefficient between the returns on any pair of assets is 1/2 and there is also a risk free asset. Assume the CAPM model is satisfied.

(i) Calculate the expected rate of return and standard deviation of the market portfolio.

(ii) Calculate the betas of the three assets.

(iii) Use solution to (ii) to find the beta of the market portfolio.

(iv) What is the risk-free rate of return implied by these returns?

(v) Describe how this model could be used to price a new asset, D.

Exercise 73 *Exercise to show that in regression of excess returns the value of the intercept must be zero. Describe why this is a test of CAPM.*

Exercise 74 *Let return on market and asset be observed.*

State

Asset

Market

(i) Find the β for the asset.

(ii) Given β in which state is it above and below the security market line?

(iii) Show that in expected terms it is on the SML.

Exercise 75 *Take two assets with betas β_A and β_B held in proportions X_A and X_B which are the market portfolio of risky assets. If the return of A is ?? and $r_f =?$, $\bar{r}_M =?$ and $\sigma_M =?$. What must be $\bar{r}_B =?$ If $\bar{r}_B =?$ what would you do? If $\bar{r}_B =?$ what would you do?*

Exercise 76 *Use the CAPM2 example for two risky assets and a simplified utility to get some cancellation.*

Exercise 77 *(i) Consider an asset with expected future price of 10 and a beta of 1.2. If* $r_f = 0.05$ *and* $\bar{r}_M = 0.1$*, what is the fair market price of the asset today?*

(ii) If the equilibrium price today is ?, what is the expected price next year?

Exercise 78 *A project costs \$1000 to undetake and its payoff is related to the market as in the table.*

State	*1*	*2*	*3*	*4*
Project				
Market				

(i) Find the return on the project in each state.

(ii) Calculate the beta of the project.

(iii) Is the PDV of the project positive or negative?

(iv) If ?? were changed to ??, would decision on project alter?

Chapter 9

Arbitrage Pricing Theory

9.1 Introduction

Arbitrage Pricing Theory (APT) is an alternative to CAPM as a theory of equilibrium in the capital market. It works under much weaker assumptions. Basically, all that is required is that the returns on assets are linearly related to a set of indices and that investors succeed in finding all profitable opportunities.

CAPM is based on very specific assumptions about the behavior of investors. It reaches clear conclusions but at the cost or restrictive assumptions. APT attempts to reach similarly informative conclusions but on the basis of weaker assumptions.

APT starts with assumptions about the distribution of asset returns and relies on approximate arbitrage arguments. It is based on the assumption that asset returns are generated by a set of underlying factors. Thus, the multi-factor model of Chapter 11 is assumed to apply exactly. The equilibrium is then obtained by asserting that there can be no unrealized returns. This results from investors arbitraging away all possible excess profits.

9.2 Returns Process

The foundation of the APT is the assumption that the return on asset i is generated by an underlying set of factors. Typical factors could include interest rates and inflation. In principle, these factors can be anything that affect asset prices and returns.

To introduce the model in its simplest form, it is assumed initially that there are only two factors. The extension of the argument to many factors will be given later.

Denote the value of factor k by F_k. APT assumes that the returns on assets are determined by unexpected changes in the factors. These unexpected changes are called "news". It is unexpected changes that matter because any information that is already expected will have been factored into returns when it first became known. The news in factor k is given by

$$f_k = F_k - E[F_k]. \tag{9.1}$$

Example 105 *Let factor* 1 *be the inflation rate. Inflation is expected to be 2% ($E[F_k] = 2$) . Inflation is announced to be 4% ($F_k = 4$). The news is 2% ($f_k = 2$).*

If there are two factors the return on asset i in APT is given by

$$\tilde{r}_i = a_i + b_{i1}\tilde{f}_1 + b_{i2}\tilde{f}_2 + \tilde{e}_i, \tag{9.2}$$

where a $\tilde{}$ denotes a random variable. In this formulation the factor loading b_{ik} measures how sensitive the return on asset i is to the news on factor k. The term $\tilde{e}_i$ is the idiosyncratic error on asset i. This is the variation in return that is unrelated to the two factors. It is assumed that

$$E\left[\tilde{f}_k\right] = 0, \tag{9.3}$$

which reflects the interpretation of $\tilde{f}_k$ as news,

$$E[\tilde{e}_i] = 0, \tag{9.4}$$

so that no systematic error is made, and, for any pair of assets i and j,

$$E[\tilde{e}_i\tilde{e}_j] = 0, i \neq j. \tag{9.5}$$

Condition (9.5) states that the non-systematic errors are uncorrelated between any two assets. This implies that the part of the variation in return on asset i that is not explained by the factors is caused by factors unique to the asset.

It is now assumed that the portfolio of each investor is well-diversified so that idiosyncratic risk can be ignored. See Section 7.6 for the justification of this approach. Only the systematic risk caused by the variation of the factors is then relevant for portfolio risk. The return on a portfolio composed of N assets with the return process in (9.2) and portfolio weights $\{X_1, ..., X_N\}$ is

$$\tilde{r}_p = a_p + b_{p1}\tilde{f}_1 + b_{p2}\tilde{f}_2, \tag{9.6}$$

where

$$a_p = \sum_{i=1}^{N} X_i a_i, \tag{9.7}$$

and

$$b_{pk} = \sum_{i=1}^{N} X_i b_{ik}, \quad k = 1, 2. \tag{9.8}$$

There are two properties of this model that need to be noted.

First, a portfolio not exposed to any risk must offer the risk-free return. So if portfolio p_0 has $b_{p_0 1} = b_{p_0 2} = 0$ then $\bar{r}_{p_0} = r_f$. This implies that

$$a_{p_0} = r_f. \tag{9.9}$$

Second, it is possible to construct a portfolio that is exposed only to the risk from a single factor. If portfolio p_k has this property and is exposed only to risk from factor k then

$$\tilde{r}_{p_k} = a_{p_k} + b_{p_k k}\bar{\tilde{f}}_k. \tag{9.10}$$

Such a portfolio can be constructed from any two distinct well-diversified portfolios. Consider two such portfolios described by

$$\tilde{r}_1 = a_1 + b_{11}\tilde{f}_1 + b_{12}\tilde{f}_2, \tag{9.11}$$

and

$$\tilde{r}_2 = a_2 + b_{21}\tilde{f}_1 + b_{22}\tilde{f}_2, \tag{9.12}$$

where $b_{11} \neq b_{21}$, $b_{12} \neq b_{22}$, and $b_{ik} \neq 0$ all i, k. Then there always exists factor proportions $\hat{X}_1$ and $\hat{X}_2$ such that

$$\hat{X}_1 b_{12} + \hat{X}_2 b_{22} = 0. \tag{9.13}$$

To see this, observe that

$$\begin{aligned}\hat{X}_1 b_{12} + \hat{X}_2 b_{22} &= \hat{X}_1 b_{12} + \left[1 - \hat{X}_1\right] b_{22} \\ &= \hat{X}_1 \left[b_{12} - b_{22}\right] + b_{22}. \end{aligned} \tag{9.14}$$

This is zero when

$$\hat{X}_1 = \frac{b_{22}}{b_{22} - b_{22}}. \tag{9.15}$$

The portfolio, p_1, with these weights has return

$$\tilde{r}_{q'} = a_{p_1} + b_{p_1 1}\tilde{f}_1. \tag{9.16}$$

Similarly if portfolio p_2 is defined by

$$\hat{X}_1 = \frac{b_{21}}{b_{21} - b_{11}}, \tag{9.17}$$

then

$$\tilde{r}_{q''} = a_{p_2} + b_{p_2 2}\tilde{f}_2. \tag{9.18}$$

Example 106 *Consider two portfolios defined by the returns processes*

$$\begin{aligned}\tilde{r}_1 &= 0.2 + 1.2\tilde{f}_1 + 0.8\tilde{f}_2, \\ \tilde{r}_2 &= 0.4 + 1.8\tilde{f}_1 + 0.6\tilde{f}_2.\end{aligned}$$

Setting $\hat{X}_1 = \frac{0.6}{0.6-0.8} = -3$ *and* $\hat{X}_2 = 4$ *gives a portfolio with return*

$$\begin{aligned}\tilde{r}_{q'} &= [-3(0.2)+4(0.4)] + [-3(1.2)+4(1.8)]\tilde{f}_1 + [-3(0.8)+4(0.6)]\tilde{f}_2 \\ &= 1+3.6\tilde{f}_1.\end{aligned}$$

Alternatively, setting $\hat{X}_1 = \frac{1.8}{1.8-1.2} = 3$ *and* $\hat{X}_2 = -2$ *gives a portfolio with return*

$$\begin{aligned}\tilde{r}_{q''} &= [3(0.2)-2(0.4)] + [3(1.2)-2(1.8)]\tilde{f}_1 + [3(0.8)-2(0.6)]\tilde{f}_2 \\ &= -0.2+1.2\tilde{f}_1.\end{aligned}$$

A portfolio, p_k, with a risk factor of $b_{p_k k} = 1$ for factor k, and zero for other factors is called a *factor portfolio* for factor k. The return on a factor portfolio for factor k is

$$\tilde{r}_{p_k} = a_{pk} + \tilde{f}_k. \tag{9.19}$$

The expected return on the factor portfolio is

$$\bar{r}_{f_k} \equiv E[\tilde{r}_{pk}] = a_{pk} + E\left[\tilde{f}_k\right] = a_{pk}. \tag{9.20}$$

The *premium* for factor k is defined by $\bar{r}_{f_k} - r_f$. This is the return over the risk-free rate for holding the factor portfolio.

Example 107 *Consider portfolio* p_1 *from the previous example. This has return generated by* $\tilde{r}_{p_1} = 1 + 3.6\tilde{f}_1$*. Consider combining this portfolio with the risk-free asset in proportions* $X_{p_1} = \frac{10}{36}$*,* $X_f = \frac{26}{36}$*. The return on this portfolio,* q*, is*

$$\begin{aligned}\tilde{r}_q &= \frac{10}{36}\left[1+3.6\tilde{f}_1\right] + \frac{26}{36}r_f \\ &= \frac{10}{36} + \frac{26}{36}r_f + \tilde{f}_1.\end{aligned}$$

If the risk-free rate is $r_f = 2$ *then*

$$E[\tilde{r}_q] = \frac{62}{36}.$$

The premium on factor 1 *is*

$$E[\tilde{r}_q] = \frac{62}{36} - 2 = -\frac{10}{36}$$

9.3 Arbitrage

Arbitrage ensures that there are no risk-free profits to be earned in equilibrium. The conclusion of APT is that this results in assets returns having a very particular structure.

Take a well-diversified portfolio q, two factor portfolios p_1 and p_2, and a risk-free asset. Here the phrase well-diversified means implies that all idiosyncratic risk has been diversified away. The argument now determines the relationship that must hold for there to be no arbitrage opportunities involving these portfolios and the risk-free asset.

Let portfolio q have factor loadings b_{q1} and b_{q2}. The return on this portfolio is

$$\tilde{r}_q = a_q + b_{q1}\tilde{f}_1 + b_{q2}\tilde{f}_2. \tag{9.21}$$

The returns on the two factor portfolios p_1 and p_2 are

$$\tilde{r}_{p_1} = a_{p_1} + b_{p_1 1}\tilde{f}_1, \tag{9.22}$$

and

$$\tilde{r}_{p_2} = a_{p_2} + b_{p_2 2}\tilde{f}_2. \tag{9.23}$$

Consider forming a portfolio, P, of the two factor portfolios and the risk-free asset. Choose the portfolio weights so that $X_{p_1} = b_{q1}$, $X_{p_2} = b_{q2}$, and $X_f = 1 - b_{q1} - b_{q2}$. The return on this portfolio is

$$\begin{aligned} \tilde{r}_P &= X_f r_f + X_{p_1}\tilde{r}_{p_1} + X_{p_2}\tilde{r}_{p_2} \\ &= [1 - b_{q1} - b_{q2}]\, r_f + b_{q1}a_{p_1} + b_{q2}a_{p_2} + b_{q1}\tilde{f}_1 + b_{q2}\tilde{f}_2 \\ &= a_P + b_{q1}\tilde{f}_1 + b_{q2}\tilde{f}_2, \end{aligned} \tag{9.24}$$

where

$$a_P = [1 - b_{q1} - b_{q2}]\, r_f + b_{q1}a_{p_1} + b_{q2}a_{p_2}. \tag{9.25}$$

Now notice that this portfolio has the same factor loadings as portfolio q. If it does not pay the same return as portfolio q then it is possible to undertake riskless arbitrage. To see this, assume that $E[\tilde{r}_q] > E[\tilde{r}_P]$. Form an arbitrage portfolio, a, with $X_q > 0$ and $X_P = -X_q < 0$. Then the return on the arbitrage portfolio, a, is

$$\begin{aligned} \tilde{r}_a &= X_q\tilde{r}_q + X_P\tilde{r}_P \\ &= X_q[\tilde{r}_q - \tilde{r}_P] \\ &= X_q\left[a_q + b_{q1}\tilde{f}_1 + b_{q2}\tilde{f}_2 - a_P - b_{q1}\tilde{f}_1 - b_{q2}\tilde{f}_1\right] \\ &= X_q[a_q - a_P] > 0, \end{aligned} \tag{9.26}$$

where the final inequality follows from the facts that $E[\tilde{r}_q] = a_q$ and $E[\tilde{r}_P] = a_P$, and $E[\tilde{r}_q] > E[\tilde{r}_P]$. This establishes that the arbitrage portfolio provides a guaranteed return for no net investment. This cannot occur in equilibrium.

Alternatively, if $E[\tilde{r}_q] < E[\tilde{r}_P]$ form an arbitrage portfolio, a, with $X_q < 0$ and $X_P = -X_q > 0$. Then the return on the arbitrage portfolio is

$$\begin{aligned} \tilde{r}_a &= X_q\tilde{r}_q + X_P\tilde{r}_P \\ &= X_P[\tilde{r}_P - \tilde{r}_q] \\ &= X_P\left[a_P + b_{q1}\tilde{f}_1 + b_{q2}\tilde{f}_2 - a_q - b_{q1}\tilde{f}_1 - b_{q2}\tilde{f}_2\right] \\ &= X_P[a_P - a_q] > 0. \end{aligned} \tag{9.27}$$

Under the assumed condition the arbitrage portfolio again provides a guaranteed return for no net investment, so this cannot be an equilibrium outcome.

The condition that must be met to ensure that there can be no arbitrage is

$$E\left[\tilde{r}_q\right] = E\left[\tilde{r}_P\right]. \tag{9.28}$$

For this condition to hold requires

$$\begin{aligned} a_q &= a_P \\ &= \left[1 - b_{q1} - b_{q2}\right] r_f + b_{q1} a_{p_1} + b_{q2} a_{p_2} \end{aligned} \tag{9.29}$$

From the property of a factor portfolio described in (9.20)

$$a_{p_1} = \bar{r}_{f_1}, \tag{9.30}$$

and

$$a_{p_2} = \bar{r}_{f_2}. \tag{9.31}$$

Using these results

$$\begin{aligned} a_q &= \left[1 - b_{q1} - b_{q2}\right] r_f + b_{q1}\bar{r}_{f_1} + b_{q2}\bar{r}_{f_2} \\ &= r_f + b_{q1}\left[\bar{r}_{f_1} - r_f\right] + b_{q2}\left[\bar{r}_{f_2} - r_f\right], \end{aligned} \tag{9.32}$$

and hence

$$\begin{aligned} \bar{r}_q &= E\left[a_q + b_{q1}\tilde{f}_1 + b_{q2}\tilde{f}_2\right] \\ &= r_f + b_{q1}\left[\bar{r}_{f_1} - r_f\right] + b_{q2}\left[\bar{r}_{f_2} - r_f\right]. \end{aligned} \tag{9.33}$$

This is the only value of expected return for portfolio q that is consistent with the absence of arbitrage opportunities. This result is a special case of the central conclusion of APT.

Example 108 *Consider three well-diversified portfolios A, B, and C with returns and factor sensitivities in the following table.*

Portfolio	*Expected return %*	b_{i1}	b_{i2}
A	*13*	*0.7*	*1.1*
B	*15*	*0.6*	*1.4*
C	*11*	*1.0*	*0.7*

Let there be a further well-diversified portfolio, D, with an expected return of $\bar{r}_D = 14\%$ *and factor sensitivities of* $b_{D1} = 0.8$ *and* $b_{D2} = 1.0$*. A portfolio, E, formed as a combination of portfolios A, B, and C will match the factor sensitivities of portfolio D if the weights* X_A, X_B, *and* X_C *are such that*

$$\begin{aligned} X_A b_{A1} + X_B b_{B1} + X_C b_{C1} &= 0.8, \\ X_A b_{A2} + X_B b_{B2} + X_C b_{C2} &= 1.0, \end{aligned}$$

and

$$X_A + X_B + X_C = 1.$$

Solving these equations gives

$$X_A = 0.4,\ X_B = 0.2,\ X_C = 0.4.$$

These weights imply that the expected return on portfolio E is

$$\overline{r}_E = 0.4(13) + 0.2(15) + 0.4(11) = 12.6.$$

Arbitrage can then be conducted between portfolio E and portfolio D to realize a return with no risk. Let the portfolio weights be X_D and X_E with $X_D > 0$ and $X_E = -X_D$. The factor sensitivities of the arbitrage portfolio are given by

$$\begin{aligned} b_{ak} &= X_D b_{Dk} + X_E b_{Ek} \\ &= X_D \left[b_{Dk} - b_{Ek} \right] \\ &= 0, \quad k = 1, 2, \end{aligned}$$

since the portfolios have the same sensitivities. Hence the systematic risk of the arbitrage portfolio is zero. The expected return on the arbitrage portfolio is

$$\overline{r}_a = X_D \overline{r}_D + X_E \overline{r}_E = X_D \left[\overline{r}_D - \overline{r}_E \right] = 1.4 X_D.$$

In principal, the return on this arbitrage portfolio can be increased without limit as X_D is raised. Therefore a positive expected return is realized without any net investment on the part of the investor. This situation cannot exist in equilibrium. In fact, as investors buy the arbitrage portfolio the return on portfolio D will be driven down and that on E driven up.

The calculations have been undertaken so far for just two factors. The conclusion obtained can be generalized immediately to any number of factors. Doing this allows the central conclusion of APT to be stated as follows.

Theorem 2 *Assume there are K factors. There will be no arbitrage opportunities only if the expected return for every asset i satisfies*

$$\overline{r}_i \simeq r_f + b_{i1} \left[\overline{r}_{f_1} - r_f \right] + ... + b_{iK} \left[\overline{r}_{f_K} - r_f \right], \tag{9.34}$$

where b_{ik} is the factor loading for asset i on factor k.

Two comments can be made. First, it can be seen that this is a direct extension of (9.33). APT predicts that the expected return on each asset is determined by the risk-free rate plus sum of each factor loading multiplied by the factor premium. Second, the result is presented as an approximation. Why has the equality in (9.33) become an approximation in (9.34)? The answer is that the exact construction leading to (9.33) applies only to well-diversified portfolios with no idiosyncratic risk. The same argument cannot be applied exactly to individual assets that do have idiosyncratic risk. However, the result holds as an approximation for the following reason: if there are many assets with the same factor risks but different idiosyncratic risks it is possible to form a well-diversified portfolio of these assets that has only factor risk. This portfolio will satisfy the relation exactly. If individual assets are too far away from satisfying it, then it is possible to approximately arbitrage against them. This ensures all assets approximately satisfy the APT relation.

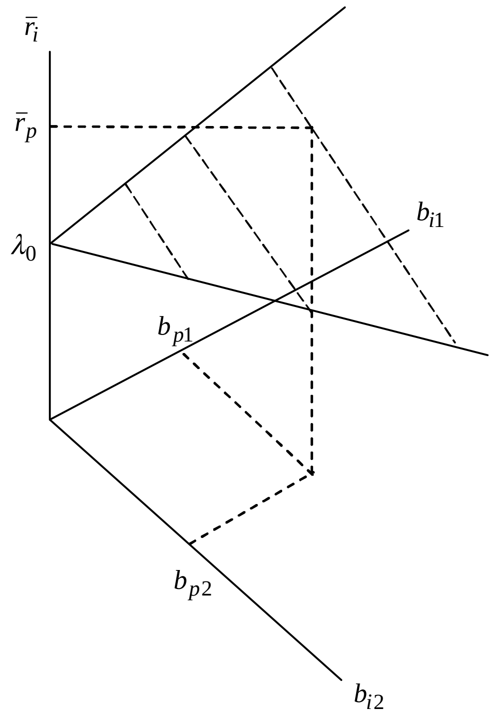


Figure 9.1: The APT equilibrium plane

9.4 Portfolio Plane

Further insight into the implications of APT can be obtained by looking at the geometrical properties of the equilibrium. This will also show what is happening behind the arbitrage argument.

Return to the two factor case of APT. The expected return on a well-diversified portfolio p can be written as

$$\begin{aligned}\bar{r}_p &= r_f + b_{p1}\left[\bar{r}_{f_1} - r_f\right] + b_{p2}\left[\bar{r}_{f_2} - r_f\right] \\ &= \lambda_0 + \lambda_1 b_{p1} + \lambda_2 b_{p2}, \end{aligned} \tag{9.35}$$

where $\lambda_0 \equiv r_f$, $\lambda_1 \equiv \left[\bar{r}_{f_1} - r_f\right]$, and $\lambda_2 \equiv \left[\bar{r}_{f_2} - r_f\right]$.

Figure 9.1 displays a three-dimensional graph with factor sensitivities, b_{i1} and b_{i2}, on the two horizontal axes and expected return, $\bar{r}_i$, on the vertical axis. Then the APT condition says that a portfolio can be represented by a point on a plane in the three dimensional space. This plane intercepts the vertical axis at λ_0. It has gradient λ_1 in the direction of b_{i1}, and gradient λ_2 in the direction of b_{i2}. The location of portfolio p is shown as a point in the plane. APT predicts that in equilibrium every well-diversified portfolio must be located somewhere on the plane, and that individual assets must be close to the plane.

This argument can be viewed another way. Assume that there are three well-diversified portfolios, A, B, and C, and two factors. Also assume that the factor sensitivities, b_{ik}, and the expected returns, $\overline{r}_i$, are known. Equilibrium for the APT then implies that there must be values of λ_0, λ_1, and λ_2 that satisfy

$$\overline{r}_A = \lambda_0 + \lambda_1 b_{A1} + \lambda_2 b_{A2}, \quad (9.36)$$

$$\overline{r}_B = \lambda_0 + \lambda_1 b_{B1} + \lambda_2 b_{B2}, \quad (9.37)$$

$$\overline{r}_C = \lambda_0 + \lambda_1 b_{C1} + \lambda_2 b_{C2}. \quad (9.38)$$

The values of λ_0, λ_1, and λ_2 can be found by solving this system of equations. The system can be written in matrix notation as

$$\begin{bmatrix} \overline{r}_A \\ \overline{r}_B \\ \overline{r}_C \end{bmatrix} = \begin{bmatrix} 1 & b_{A1} & b_{A2} \\ 1 & b_{B1} & b_{B2} \\ 1 & b_{C1} & b_{C2} \end{bmatrix} \begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \end{bmatrix}, \quad (9.39)$$

so that

$$\begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 1 & b_{A1} & b_{A2} \\ 1 & b_{B1} & b_{B2} \\ 1 & b_{C1} & b_{C2} \end{bmatrix}^{-1} \begin{bmatrix} \overline{r}_A \\ \overline{r}_B \\ \overline{r}_C \end{bmatrix}. \quad (9.40)$$

The inverse of the matrix will exist when the three assets have sufficiently different factor loadings.

Example 109 *Using the data in C, the matrix system for the returns on assets A, B, and C is*

$$\begin{bmatrix} 13 \\ 15 \\ 11 \end{bmatrix} = \begin{bmatrix} 1 & 0.7 & 1.1 \\ 1 & 0.6 & 1.4 \\ 1 & 1.0 & 0.7 \end{bmatrix} \begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \end{bmatrix}.$$

Inverting the system gives

$$\begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 19.6 & -12.2 & -6.4 \\ -14.0 & 8.0 & 6.0 \\ -8.0 & 6.0 & 2.0 \end{bmatrix} \begin{bmatrix} 13 \\ 15 \\ 11 \end{bmatrix}.$$

So the coefficients can be solved as

$$\begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 1.4 \\ 4.0 \\ 8.0 \end{bmatrix}$$

The APT then assets that for every well-diversified portfolio, p, it must be the case that

$$\overline{r}_p = 1.4 + 4b_{p1} + 8b_{p2}.$$

The APT plane can be used to illustrate how arbitrage can arise. Consider a well-diversified portfolio, q, that is located at a point below plane. By combining well-diversified portfolios that lie on the plane it is always possible to construct

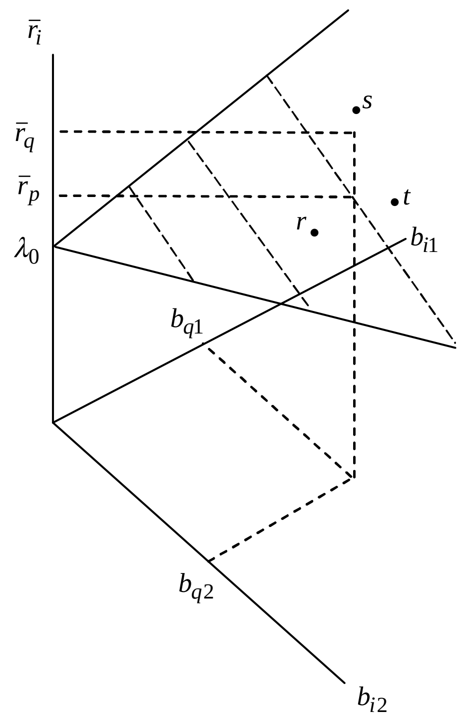


Figure 9.2: An arbitrage opportunity

a portfolio, p, that lies on the plane directly above the location of q. Since p lies directly above q it has the same factor loadings but a higher expected return. An arbitrage portfolio can then be formed by buying p and selling q. The converse argument works if there is a well-diversified portfolio that lies above the plane.

This argument is illustrated in Figure 9.2. Portfolio q lies above the plane. A portfolio of r, s, and t is formed to produce p that lies directly below q. The arbitrage portfolio then consists of buying q and selling p.

Example 110 *The arbitrage argument used in Example 108 can be reconsidered from the perspective of the APT plane. For portfolio D*

$$\begin{aligned} \bar{r}_D &> 1.4 + 4b_{D1} + 8b_{D2} \\ &= 1.4 + 4(0.8) + 8(1.0) \\ &= 12.6. \end{aligned}$$

Consequently, portfolio D lies above the APT plane determined by portfolios A, B, and C. In contrast, portfolio E lies on the plane since

$$\bar{r}_E = 1.4 + 4b_{1E} + 8b_{2E} = 12.6.$$

Portfolio E lies directly below portfolio D since it has the same factor loadings. It is this property that permits the arbitrage portfolio to be constructed.

9.5 General Case

It is now possible to state the general conclusion of APT. The argument has already been made that if an arbitrage portfolio can be found, then investors will find it. Investment in the arbitrage portfolio will ensure that any portfolios off the plane determined by the returns of the other portfolios will be driven onto it as prices (and hence returns) change.

This argument can be formalized as follows. Let the number of factors be K. Take a set of portfolios equal in number to the number of factors. These will define a plane in $K+1$–dimensional space. Any distinct set of K portfolios will do for this purpose. For example, in the examples above an APT plane could have been constructed from portfolios A, B and D and then an arbitrage portfolio constructed against portfolio C. All that matters for the argument is that the set of four portfolios - A, B, C, D - do not lie on the same plane.

This arbitrage activity will ensure that in equilibrium there can be no portfolios either below or above the APT plane. Thus, all portfolio returns must be related by the equation of the plane relating factor loadings to expected return. The contribution of APT is to conclude that this equilibrium plane exists and to characterize its structure.

The construction of the plane is now extended to incorporate any number of factors. Take K factors with the return process for a well-diversified portfolio given by

$$\tilde{r}_i = a_i + \sum_{k=1}^{K} b_{ik}\tilde{f}_k. \tag{9.41}$$

Assume that the number of well-diversified portfolios is at least as large as the number of factors ($N \geq K$) and that the factor loadings of the portfolios are sufficiently distinct. These portfolios determine a plane in n-dimensional space which has equation for the expected return

$$\overline{r}_i = \lambda_0 + \sum_{k=1}^{n} \lambda_k b_{ik}. \tag{9.42}$$

The coefficients $\lambda_0, ..., \lambda_n$ can be found from solving

$$\begin{bmatrix} \lambda_0 \\ \vdots \\ \lambda_K \end{bmatrix} = \begin{bmatrix} 1 & b_{11} & \cdots & b_{1K} \\ \vdots & & & \vdots \\ 1 & b_{N1} & \cdots & b_{NK} \end{bmatrix}^{-1} \begin{bmatrix} \overline{r}_A \\ \vdots \\ \overline{r}_N \end{bmatrix}. \tag{9.43}$$

The arbitrage argument is that if there is a portfolio that is not on this plane, then an arbitrage portfolio can be constructed. If the factor sensitivities of the portfolio not on the plane are b_{jk}, $k = 1, ..., n$ then the arbitrage portfolio is defined by

$$\sum_{i=1}^{N} X_i b_{ik} = b_{jk}, k = 1, ..., n, \quad (9.44)$$

$$\sum_{i=1}^{N} X_i = 0. \quad (9.45)$$

APT also makes statements about the price of risk. The coefficient λ_i is the price of risk associated with factor i. That is, a unit increase in the factor loading b_{ik} will be rewarded with an increase in expected return equal to λ_i. This is just a reflection again of the fact that an investor will only accept greater variability (measured by a higher value of b_{ki}) if more return is gained. In equilibrium, the λ_is determine just how much greater this risk has to be.

9.6 Implementation of APT

The practical implementation of the APT involves three steps. These steps are:

- Identification of the factors
- Estimation of the factor loadings for each asset
- Estimation of the factor premia

These steps are now considered in turn.

The first step is the identification of the factors. Recall that in CAPM the single factor was uniquely identified as the market portfolio. In contrast, the APT model does not itself specify what the factors are. This means the factor have to be determined empirically. This process of identifying the factors should be guided by consideration of what forms of news might correlate assets prices.

One approach is to consider various macroeconomic factors. These would be tested to see if they are significant for returns. The macroeconomic factors that are typically considered include:

- Changes in GDP growth
- Changes in Treasury bill yields (inflation proxy)
- Changes in spread between Treasury bill and Treasury bonds (predictor of interest rate changes)
- Changes in risk premium on corporate bonds
- Changes in oil prices

An alternative approach is to use the statistical technique of factor analysis to extract the factors from data on the returns of assets. This involves extracting factors from the covariance matrix.

The second step is to obtain the factor loadings by regressing past asset returns on the factors.

Factor premia are found in the third step by constructing factor portfolios.

The APT theorem can then be used to price assets given the identified factors, factor loadings, and factor premia.

Example 111 *The standard example is drawn from Fama and French. The factors are*

1. Market factor: return on market index minus its mean

2. Size factor: return on small stocks minus return on large stocks

3. Book-to-market factor: return on high book-to-market stocks minus return on low book-to-market stocks

The factor premia are

Factor	*1. Market*	*2. Size*	*3. Book-to-market*
Premium	*5.2*	*3.2*	*5.4*

Hence the APT equation for an asset is

$$\bar{r}_i = r_f + b_{i1}\,[5.2] + b_{i2}\,[3.2] + b_{i3}\,[5.4]$$

The results estimated for a range of sectors are in the table.

	Three-factor APT				*CAPM*
	b_{i1}	b_{i2}	b_{i3}	*Premium*	*Premium*
Aircraft	*1.15*	*0.51*	*0.00*	*7.54*	*6.43*
Banks	*1.12*	*0.13*	*0.35*	*8.08*	*5.55*
Chemicals	*1.13*	*-0.03*	*0.17*	*6.58*	*5.57*
Computers	*0.90*	*0.17*	*-0.47*	*2.46*	*5.29*
Construction	*1.21*	*0.21*	*-0.09*	*6.42*	*6.52*
Food	*0.88*	*-0.07*	*-0.03*	*4.09*	*4.44*
Petroleum and gas	*0.95*	*-0.35*	*0.21*	*4.93*	*4.32*
Pharmaceuticals	*0.84*	*-0.25*	*-0.63*	*0.09*	*4.71*
Tobacco	*0.86*	*-0.04*	*0.24*	*5.56*	*4.08*
Utilities	*0.79*	*-0.20*	*0.38*	*5.41*	*3.39*

9.7 APT and CAPM

APT, and the multi-factor model, are not necessarily inconsistent with CAPM. In the simplest case with one factor, the two are clearly identical. With more than factor further conditions must be met for these to be identical.

APT is based on factors and arbitrage arguments. CAPM is based on optimal portfolios and demand and supply. These are not fundamentally different processes so it is not surprising that the two models produce similar predictions. It is the additional assumption that all investors construct the same efficient frontier that gives additional precision to the predictions of CAPM.

To obtain an insight into the conditions under which APT and CAPM are identical, assume that returns are generated from two factors so

$$\tilde{r}_i = a_i + b_{i1}\tilde{f}_1 + b_{i2}\tilde{f}_2. \tag{9.46}$$

The equilibrium from the APT model is then determined by the equilibrium equation

$$\overline{r}_i = r_f + \lambda_1 b_{i1} + \lambda_2 b_{i2}, \tag{9.47}$$

where the condition $\lambda_0 = r_f$ has been used. The interpretation of λ_k is that this is the return above the risk-free rate earned by an asset with $b_{ik} = 1$ and all other values of $b_{ij} = 0$. From the CAPM the value of this excess return should be

$$\lambda_i = \beta_{\lambda_i}\left[\overline{r}_M - r_f\right]. \tag{9.48}$$

Substituting this into (9.47) gives

$$\begin{aligned}\overline{r}_i &= r_f + b_{i1}\beta_{\lambda_1}\left[\overline{r}_M - r_f\right] + b_{i2}\beta_{\lambda_2}\left[\overline{r}_M - r_f\right] \\ &= r_f + \left[b_{i1}\beta_{\lambda_1} + b_{i2}\beta_{\lambda_2}\right]\left[\overline{r}_M - r_f\right].\end{aligned} \tag{9.49}$$

This is exactly the CAPM model where $\beta_i = b_{i1}\beta_{\lambda_1} + b_{i2}\beta_{\lambda_2}$. The two remain consistent provided this identity holds.

9.8 Conclusions

APT in practice gives a reasonable description of risk and return

Factors used in applications seem plausible

No need to construct a market portfolio or to measure return on this

But

Model does not identify the factors

Factors might change over time

Multi-factors models require more data

Exercise 79 *find something*

Exercise 80 *add something*

Exercise 81 *and something*

Chapter 10

Empirical Testing

10.1 Introduction

The equilibrium models of the previous two chapters make a number of predictions about the structure of asset returns. These predictions can be very helpful in pricing assets, but only have validity if the underlying model is correct. This raises the question of how the predictions of the models can be tested.

The tests of these equilibrium models are important since they influence how the market is viewed. If either is correct, then that gives a direct influence upon how investment decisions are made and evaluated. For instance, it CAPM is true then it is unnecessary to purchase anything but the market portfolio. Alternatively, if APT is true then an alternative process is implied for the pricing of assets.

It might be thought that the models can be tested by considering the realism of the underlying assumptions. If this were the case the CAPM would have received very little attention since the strictness of its assumptions would have seen it dismissed very quickly. However, an argument can be made that a model should be judged by its predictions and not by its assumptions. Since the predictions of the CAPM are so clear, precise, and useful this arguments provides strong support for testing it carefully despite its doubtful assumptions.

This chapter will look at the testing of the CAPM. It will describe the development of the testing methodology and the current position concerning the assessment of the model.

The chapter will also look at the implementation of APT.

10.2 Testing CAPM

The two forms of the CAPM will first be reviewed. Then the early tests will be considered. Next will come the literature on anomalies. Then a summary of Roll's critique.

10.2.1 Forms of CAPM

Chapter 8 has described the Sharpe-Lintner and Black forms of the CAPM. These are briefly reviewed to provide the basis for testing.

The Sharpe-Lintner form of the CAPM assumes that lending and borrowing are possible at the riskfree rate. For this version of the model the security market line is given by

$$E(r_i) = r_f + \beta_i^M \left[E(r_M) - r_f \right], \tag{10.1}$$

where β_i^M is defined by

$$\beta_i^M = \frac{cov\,(r_i, r_M)}{var(r_M)}. \tag{10.2a}$$

The security market line is often expressed in terms of returns in excess of the riskfree rate. Define the excess return on asset i by $z_i \equiv r_i - r_f$. The security market line can then be written in terms of excess returns as

$$E(z_i) = \beta_i^M E(z_M), \tag{10.3}$$

with

$$\beta_i^M = \frac{cov\,(z_i, z_M)}{var(z_M)}. \tag{10.4}$$

If the riskfree rate is non-stochastic then the two forms of β_i^M in (10.2a) and (10.4) are identical. When applied in empirical work a proxy for the riskfree rate will be stochastic, so the two β_i^M may differ. Most empirical work uses the formulation in terms of excess returns.

To provide a testable theory the equations is converted to its empirical counterpart. The security market line is written in terms of ex ante expected returns. In contrast, historical data provide information on ex post realized returns. Even if the model were exactly true the ex post data would never exactly satisfy the (10.3) since the ex post returns are realizations of random variables. This justifies writing the empirical counterpart of (10.3) as

$$z_{i,t} = \alpha_i + \beta_i z_{M,t} + \varepsilon_{i,t}, \tag{10.5}$$

where $z_{i,t}$ is the excess return on asset i at time t, $z_{M,t}$ is the excess return on the market at time t, and $\varepsilon_{i,t}$ is the error at time t. If the CAPM is true then on average the error, $\varepsilon_{i,t}$, will have a mean value of zero.

The empirical test have focused on three implications of (10.3) for the empirical model. These are:

- The estimated intercept $\hat{\alpha}_i$ should be zero;
- Beta captures all cross-sectional variation so no other variable should be significant in the regression;
- The market risk premium should be positive, implying $\hat{\beta}_i > 0$.

The Black version of the CAPM applies when borrowing and lending are not possible. In this version of the CAPM the security market line is constructed using the excess return on the market portfolio relative to the zero beta portfolio associated with the market. The zero beta portfolio is the portfolio with minimum variance of all portfolios uncorrelated with the market portfolio.

The security line for the Black CAPM can be written as

$$E(r_i) = E(r_{0M}) + \beta_i^M \left[E(r_M) - E(r_{0M})\right], \tag{10.6}$$

where r_{0M} is the return on the zero-beta portfolio. In applications the zero-beta portfolio return is treated as an unobserved quantity.

The Black model can be tested as a restriction on the market model. For the market model

$$E(r_i) = \alpha_i^M + \beta_i^M E(r_M). \tag{10.7}$$

If the Black model is true it implies

$$\alpha_i^M = E(r_{0M}) \left[1 - \beta_i^M\right], \tag{10.8}$$

for every asset i. So, the Black models restricts the intercept of the market model.

In both cases the basic prediction of the CAPM is that market portfolio lies on the efficient, so is itself an efficient portfolio. This requirement can provide a direct test of CAPM. The security market line is then a further set of predictions based on the efficiency of the market portfolio. This constitutes a secondary form of test.

10.2.2 Initial Testing

Initial testing of the CAPM considered the security market line. The technique used was a two-pass regression. In the first pass the betas of individual securities or of portfolios were estimated. This was a time series regression of asset returns on market returns. The benefit of using a portfolios is that this raises the accuracy of the estimated betas. The second pass regression was a cross-section regression of beta on average returns???. This should give the security market line.

Jensen (1968): The idea was to test the performance of mutual funds. Used data on returns on 115 open end mutual funds for 1955 - 1964. Used S&P 500 as the market index. The returns were continuously compounded. Ran a time series regression of excess return on portfolio on excess return on the market with an intercept included. Concluded that the majority of intercept terms were negative and that very few were statistically significantly different from zero.

Black, Jensen, and Scholes (1972): Begin by discussing the time series approach used by Jensen (1968) to test whether α is zero. Observe that this would be correct if the residuals for the different assets are independent. Claim instead that there is evidence that the residuals are correlated $E\left[\tilde{e}_{i,t},\tilde{e}_{j,t}\right]\neq 0$ which invalidates the simple test. Observe that this can be overcome by using portfolios of assets rather than different assets on the grounds that the non-independence will then enter into the intercept term. Also choose to group assets to obtain maximum dispersion in the betas of the portfolios. To avoid selection bias the assets are chosen on the basis of betas estimated on historical data that does not enter into the estimation of the security market line.

Sample is all stocks traded on NYSE from 1926 to 1966. Begin with subperiod 1926 to 1930. Compute betas for all stocks using an equally-weighted portfolio of all stock on NYSE as the market portfolio. Riskfree rate is the 30-day rate on US Treasury Bills for 1948 - 1966, and the dealer commercial paper rate from 1926 - 1947.

Rank stocks on basis of betas estimated on the basis of the five years of data from January 1926 December 1930. Construct 10 portfolios with highest beta stock in portfolio 1, and so on downwards. Next compute return on each portfolio for the 12 months of 1931. At the end of this year again compute betas for every stock from 1927 to 1931. Reform 10 portfolios. Repeat this process through to 1965.

Then estimate the alphas and betas of the 10 portfolios by using the all 35 years of monthly data on the returns on these portfolios. This is a time series analysis. The results are summarized in the table. Note the finding that the alphas for the high beta portfolios tend to be negative, and positive for the low beta portfolios. But not many of the alphas are significantly different from zero. Paper notes that there is the issue that the parameter values are not constant over sub-periods. Thus invalidating part of the testing. However, note that the Black model implies $\alpha_i^M = E(r_{0M})\left[1-\beta_i^M\right]$, so if $E(r_{0M}) > 0$ high beta portfolios will have negative alphas. This can explain the pattern in the table.

Portfolio	$\hat{\alpha}\times 10^2$	$t(\hat{\alpha})$	$\hat{\beta}$
1	-0.0829	-0.4274	1.5614
2	-0.1938	-1.9935	1.3838
3	-0.0649	-0.07597	1.2483
4	-0.0167	-0.2468	1.1625
5	-0.0543	-0.8869	1.0572
6	0.0593	0.7878	0.9229
7	0.0462	0.7050	0.8531
8	0.0812	1.1837	0.7534
9	0.11968	2.3126	0.6291
10	0.2012	1.8684	0.4992

The same data is then used to conduct a cross-section analysis. This is a regression of the excess return on the beta value. Use a Black version of the CAPM to obtain results. The gradient of the estimated security market line is 0.01081, so a market risk premium of 1.081% per month, or 12.972% per year. The intercept is 0.00519 or 0.519%, which is an annual of 6.228%. Higher than the average interest rate on riskfree bonds. This is relatively high. Interpretation depends on whether Sharpe-Lintner of Black CAPM is true. Claim this is CAPM where borrowing at this rate is not possible but lending is. So support the Black form of CAPM.

Overall, Black, Jensen, and Scholes deliver results that support the Black form of CAPM. The constructed beta explains much of the cross-sectional variation.

This approach has been repeated on newer data by Fama and French (2004?). Estimate a pre-ranking beta for every NYSE, AMEX, and NASDAQ stock in the CRSP database in December of every year using 2 to 5 years of prior monthly returns. Form ten value-weighted portfolios based on these pre-ranking betas and compute returns for the next twelve months. Repeat every year from 1928 to 2003. Give 912 monthly returns on ten beta-sorted portfolios. Plot beta against the average annualized monthly return. Plot a predicted security market line for the Sharpe-Lintner CAPM by using one-month Treasury bill rate and the average excess CRSP market return for 1928 - 2003. The figure they give (****) shows that predicted return on low beta portfolios is too high, and predicted return on high beta portfolios is too low. But the plotted line is approximately linear.

Fama and MacBeth (1973): focus on security market line but try to predict future returns using estimates from previous periods. Same data as for BJS and the same market portfolio.

Compute beta for every stock listed from 1926 to 1929. Rank by beta and form 20 portfolios. Estimate beta of each portfolio by relating monthly returns to market index from 1930 to 1934. Use these betas to predict returns in the months from 1935 to 1938. For each month relate monthly return to beta to derive a monthly security market line.

Then want to test whether the relation is linear. Also, CAPM says only beta matters. Add the value of beta squared to the security market line. This should have a zero coefficient. Further test whether it is only beta that matters by adding a residual variance term. That is, add as an explanatory variable to each security market line the average residual variance of the stocks in the portfolio. Hence estimate the regression

$$r_{p,t} = a_0 + a_1\hat{\beta}_p + a_2\hat{\beta}_p^2 + a_3 v + \varepsilon_{p,t}, \tag{10.9}$$

where v is the residual variance.

This process is then repeated. Stock betas are estimated for 1934 to 1938. Portfolios formed, portfolio betas estimated, and then the test equation run from 1939 to 1942. This gives a total of 390 estimates of a_0, to a_3. Then test the values of the means of these.

These results are summarized in the equations.

$$r_{p,t} = \underset{0.0061^*}{a_0} + \underset{0.0085^*}{a_1} \hat{\beta}_p + \varepsilon_{p,t} \tag{10.10}$$

$$r_{p,t} = \underset{0.0049^*}{a_0} + \underset{0.0105^*}{a_1} \hat{\beta}_p + \underset{-0.008}{a_2} \hat{\beta}_p^2 + \varepsilon_{p,t} \tag{10.11}$$

$$r_{p,t} = \underset{0.0020}{a_0} + \underset{0.0114^*}{a_1} \hat{\beta}_p + \underset{-0.026}{a_2} \hat{\beta}_p^2 + \underset{0.0516}{a_3} v + \varepsilon_p \tag{10.12}$$

In these the * denotes significance at the 10% level (i.e. 90% sure that the coefficient is different from 0).

Results;are broadly consistent with the theory. Higher beta means more return. Beta squared is not significant so no nonlinearity. Residual variation is not significant. But mean value of intercept is too high compared to riskfree rate. Again this can be assigned to inability to borrow at this rate. Hence the results support the Black zero-beta model rather than the Sharpe-Lintner model.

These two sets of results were both very supportive. This was the case with early tests (Black et al. 1972, Friend and Blume 1973, Fama and MacBeth 1973). But this perspective did not last for long.

10.2.3 Anomalies

Now on to the anomalies literature. An anomaly is a significant empirical relationship that the theory predicts should not be there. The CAPM model implies that the only variable that is significant for predicting the return on an asset or portfolio is beta. If any other variable is found to the significant in the relationship then there is an anomaly. The result reported in (10.12) shows what the regression should be like. Neither beta squared nor $v$are significant in the regression. However, empirical research did isolate other variables that are significant.

The size effect: The size effect relates to the significance of market valuation in the regressions. The results show that firms with low market capitalizations (small caps) earn higher return than predicted by CAPM and high cap firms earn lower returns. This result became apparent when portfolios were first sorted on beta and then on size. The abnormal return on small caps is about 2-4% per year. (Banz 1981, Fama and French 1992).

The value effect: The value effect is that valuation ratios are related to return. The most significant of these anomalies is the book-to-market anomaly. The book value of a firm is the value of its assets minus its liabilities. The market value of a company is the stock market value of its shares. The book-to-market ratio is the ratio of these two values. Value stocks are defined as having a low market value relative to firms fundamentals (high book-to-market). The anomaly is that value stocks earn a premium. (Basu 1977, Basu 1983, Jaffe et al 1989, Rosenberg 1985, Chan 1991, Fama and French 1992).

The momentum effect: The momentum effect is that losers tend to lose again and winners tend to win again. This holds in the short-run, for up to a year. (Jegadeesh 1990, Jegadeesh and Titman 1993, Fama and French 1996). Estimates for abnormal returns are 4-6% per year. The momentum effect os overturned in the long-run (DeBondt and Thaler 985, Chopra et al. 1992).

10.2.4 Roll critique

The Roll critique focuses on two aspects. What has been tested in the papers described above. And what can be tested. The point of the Roll critique is first to argue that tests such as those above will always find results in agreement with the CAPM because of their structure. Then second to argue that the core prediction of the CAPM is that the market portfolio is efficient but that this cannot be tested because of the difficulty of constructing a market portfolio.

The argument behind point one is to consider what is the outcome if the tests above are repeated but just using an entirely random process to generate the monthly asset returns. What is concluded is that following the process outlined in the texts of grouping assets into portfolios on the basis of betas then estimating the portfolio betas then regressing on returns will almost always generate a fit similar to the predicted by the CAPM. In other words, it is the process of construction that leads to the outcome and not the underlying behavior.

More particularly, there will always be a good fit for the estimated security market line if the betas are calculated relative to a market proxy that is includes all the assets under consideration and that is mean-variance efficient. This holds whether or not the true market portfolio is mean-variance efficient. So, any study finding a good fit to the security market line is showing only that the market proxy is efficient. On the other hand, if the fit is poor this is because the selected market proxy is inefficient.

The second point is that the theory is untestable. To see this argument consider the NYSE data used above. Imagine computing the efficiency frontier for the NYSE stocks and constructing the equally-weighted market portfolio of NYSE stocks. If the market portfolio is found to be inefficient (i.e. provide a return given its risk that is statistically significantly inside the frontier) can that be taken as proof CAPM is wrong? The answer has to be no: the analysis has only concluded that market portfolio for NYSE is inefficient, but this is not the overall market portfolio as enters the CAPM (there are stocks quoted on other US exchanges, the entire range of bonds, derivatives, and all international assets at least to be considered. In addition, human capital can be added to the list and other non-financial assets. Even if all these assets could be identified some of these are traded in thin markets so returns are hard to establish.). Expressed another was, the NYSE is just a sample of the total set of assets, and the CAPM does not predict that a constructed market portfolio for this sample should be efficient for that sample.

This strength of this point is enhanced by the observation (Roll and Ross 1994, Kandel and Stambaugh 1995, Grauer 1999) that using a proxy that deviates only slightly from the market portfolio (excluding some assets or using wrong weights) can cause a poor fit to the security market line even if CAPM is true, and a good fit when it is not. This is caused by the fact that small changes in portfolio weights can have large effects upon the estimated betas of the individual assets.

Thus the Roll critique is that it is not possible to really construct a market portfolio to test the CAPM. Every test that has been done uses only a restricted set of assets so is not a test since the outcome is sensitive to the market proxy. This leads to the perspective that the CAPM as such cannot be tested.

10.2.5 Further Issues

Ex ante: Remember that the CAPM is about asset demand based on the expected returns. This means that the model is correctly formulated in terms of ex ante variables. However, testing can only use ex post observations of realized returns.

If it is assumed that the ex post observations are draws from the ex ante distribution, that the distribution is invariant over the time of the sample, and that the market portfolio has not changed then the ex post data will be a statistically good sample of the ex ante distribution.

But there is much evidence that the returns distribution changes over time. One reason is the business cycle. Risk premiums are higher during recessions and lower during expansions. The relative riskiness of assets also changes over time as the fundamentals of the underlying firms change.

Hence the true market portfolio will change over time. Using a proxy that does not very over time can then generate any possible outcome with regard to the fit to the security market line. Expressed differently, the true market portfolio depends on the distribution of returns that is conditional on observed economic state. Estimating means and beta on an unconditional distribution will not constitute a test.

Fama and MacBeth (1973) try to overcome this by using a moving window. The excess returns on the benchmark portfolios are regressed on the betas estimated from the previous 60 monthly observations. There is a trade-off here between a window that is so long that it capture a period of much economic change, and the need to have enough data to have good estimates.

An alternative is to include economic indicators in the regession of beta on the risk premium. This is intended to capture the conditionality. For example, can use the T_bill rate, the credit spread, and the term spread. But there is little theory to guide the choice. Ghysels (1998) finds that these models can give greater pricing errors than unconditional models.

Finally, the length of the investment horizon also matters. Levhari and Levy (1977) show that monthly betas are not the same a annual betas. So, results will depend on the time length used. Kothari et al (1995) show that annual are better.

There are also potential statistical flaws. First, in the second-pass regression it has to be assumed that the error process is the same for all of the portfolios. There is evidence that this is not the case. Second, the estimation of the betas in the first pass is treated as independent of the estimation of the empirical security market line in the second pass. This ignores the errors from the estimation of beta when minimizing the errors from the security market line. A different result would emerge if the sum of both errors was minimized. This biases the intercept of the second-pass away from zero and biases the slope toward zero. Miller and Scholes (1972) show that this can be significant if individual assets are used. Employing well-diversified portfolios in the second pass helps to cancel some of these errors.

Data-mining: searching through data for a significant relationship. Invalidates statistical test. Black (193a, b): Fama and French (1992) book-to-market effect is due to data mining. Data-snooping: looking at past results to find direction for future research. Should really use new data. But this is a problem since data are limited. Sample-selection: data availability results in some assets being excluded from analysis. For example, survivorship bias means that only firms that have survived for several years are considered. Kothari et al (1995) argued that failing firms (not included) would have high book-to-market ratios and low returns. Omitting them biases results and may explain book-to-market. Also, delistings frequently remove the poor performing companies.

10.2.6 Recent Tests.

Shanken (1987) observes that it is possible to test whether a proxy for the market portfolio is inefficient relative to the efficient frontier for the subset of assets that are considered. What cannot be tested is whether the true market portfolio is efficient for the global frontier. What is known is that the global frontier must be outside the frontier for the subset of assets. The question is then whether the true market portfolio is inefficient relative to the frontier for the subset. What Shanken shows is that if it is assumed that expected returns are linearly related to the covariance with the true market portfolio and the correlation coefficient between the proxy and the true is approximately 0.8 then the CAPM can be rejected with 95% confidence if inefficiency can be established.

10.3 Implementing APT

What factors should be included?
Joint test of factor choice and model.

10.4 Conclusions

What are these? Models don't fit?

Exercise 82 *Prove that (10.2a) and (10.4) are identical when the riskfree rate is constant. What are the differences when it is not constant?*

Exercise 83 *add*

Exercise 84 *add*

Chapter 11

Efficient Markets and Behavioral Finance

Add some chat

11.1 Introduction

This chapter now tests a very basic feature of model.

11.2 Efficient Markets

Introduce the idea of an efficient market.

Efficient Market

"A (perfectly) efficient market is one in which every security's price equals its investment value at all times."

- if efficient, information is freely and accurately revealed

Types of Efficiency

Form of efficiency	embodied in prices
Weak	prices of securities
Semistrong	publicly available information
Strong	public and private information

Interpretation: cannot make excess profits using the form of information embodied in prices

Evidence: markets are at least weak-form efficient, strong is very doubtful

This finding is not surprising given the number of professional and amateur investors attempting to find profitable opportunities.

11.3 Tests of Market Efficiency

11.3.1 Event Studies

Look at the reaction of security prices after new information is released
- markets appear to perform well

11.3.2 Looking for Patterns

The return on an asset is composed of
- the risk free rate
- a premium for risk

But latter can only be predicted via a model so finding patterns is a joint test of model and efficiency.

- one finding here is the "January effect" (returns abnormally high)

11.3.3 Examine Performance

Do professional investors do better?
- problem of determining what is normal, again a joint test
- problem of random selection

Results

Those with inside information can always do well so strong-form is usually rejected

- e.g. trading of company directors

Tests of semi-strong often isolate strategies that earn abnormal returns but usually not enough to offset transactions costs.

Weak-form - some possibility that investors overreact to some types of information.

11.4 Market Anomolies

Is it worht listing anomolies or are these part of the section above?

11.5 Excess Volatility

Does this fit in the previous section?

11.6 Behvioral Finance

Look at this as an explanation of some of the failure of market efficiency.

11.7 Conclusion

Put some here: things that need to be explained.

Exercise 85 *put in exercise*

Exercise 86 *put in the next exercise*

Exercise 87 *add another.*

Part V

Fixed Income Securities

Chapter 12

Interest Rates and Yields

12.1 Introduction

Bonds are securities that promise to pay a fixed income and so are known as *fixed income securities*. They are important investment instruments in their own right. The returns on bonds are also important in determining the structure of interest rates on different types of loans.

The income from a bond takes the form of a regular coupon payment and the payment of principal on maturity. One central issue is to find a method of comparison of bonds that can have very different structures of payments and lengths of maturity. Although the promised payments are known at the time the bond is purchased, there is some risk of default. This provides a role for ratings agencies to assess the risk of bonds.

One special case of a bond is the *risk-free security* that has played such a prominent role in the theoretical analysis. In practice, the risk-free security is typically taken to be a United States or a United Kingdom short-term bond. These have little risk of default so that their payments are virtually guaranteed. Even these bonds are not entirely risk-free since there is always some risk due to inflation being unpredicatble.

The chapter first discusses different types of bond. Then it moves towards making comparisons between bonds. The first comparison is based on the assessment of risk characteristics as measured by rating agencies. Then bonds are compared using the concept of a yield to maturity. Following this, the focus is placed upon interest rates. Spot rates and forward rates are related to the payments made by bonds and it is shown how these interest rates are used in discounting. Finally, the chapters looks at the concept of duration, which measures a further property of a bond, and this is related to the price/yield relationship.

12.2 Interest Rate Calculations

This secton will introduce some important interest rates and then do some basic interest calculations.

12.2.1 Significant Rates

There are several important interest rates.

The Treasury rate is the interest rate earned on Treasury bills. These bills are described in more detail below. They are issued by governments to finance short-term borrowing. It is assumed there is no chance of default for treasury bills issued by the US, UK, and other major governments.

The LIBOR rate (which is the abbreviation of the London Interbank Offered Rate) is the rate at which one bank will be willing to make a large deposit with another bank. It is often used as the safe rate of interest in financial calculations since it is the lowest rate that is generally available for short-term depositis. Large banks quote these rates in most major currencies.

The Repo rates is defiend by the terms of repo (or Repossession) agreements. A repo is an agreement to sell securities now and buy them back at a higher price. The difference in prices determines the interest rate. Repos permit short-term borrowing at low rates because the securities are collateral. They allow the borrowers to obtain finance on cusotmized terms.

The US Federal Funds Rate is the rate of interest that is paid on overnight funds that banks deposit with the Federal Reserve. This interest rate is part of monetary policy. The Federal Reserve undertakes market intervention to target this rate. All other interest rates are above this rate. The Bank of England Base Rate operates in the same way.

12.2.2 Discrete Interest

Start with annual then make more frequent

The compounding frequency used for an interest rate is the unit of measurement

The difference between quarterly and annual compounding is the number of times interest is added (four versus one)

If an amount a is invested for t years at rate r compounded once per year then the final amount is

$$b = a\left(1+r\right)^{t}. \tag{12.1}$$

Example 112 *In two years time* \$100 *invested at an interest rate of* 10% *with the interest re-invested grows to* $100\left(1+0.1\right)^{2} = \$121$.

Equally, an amount b that will be received in n years is today worth

$$a = \frac{b}{\left(1+r\right)^{t}}. \tag{12.2}$$

The latter must be the case. Why? Borrow a today and payoff in t years with amount b. This is an arbitrage argument.

Example 113 *If the interest rate is* 5% *then* \$200 *that will be received in 3 years time is worth* $\frac{200}{(1+.05)^3} = \$172.77$ *today.*

Semi-annual compouding: need to develop this. Consider investing amount a at an annual interest rate r compounded twice per year (semi-annually). The first interest payment is of amount $a\frac{r}{2}$. This gives amount

$$a\left(1+\frac{r}{2}\right). \tag{12.3}$$

The next interest payment is $\frac{r}{2}$ on this amount. So the final amount after one year is

$$\left(a\left(1+\frac{r}{2}\right)\right)\left(1+\frac{r}{2}\right) = a\left(1+\frac{r}{2}\right)^2. \tag{12.4}$$

Example 114 *Consider* \$100 *invested at an annual rate of* 10% *with semi-annual compounding. After six months the first interest payment is* \$5. *At the end of the year the second interest payment is* $\$105 \times \frac{0.10}{2} = \$5.25$. *The final value after the two interest payments is* \$110.25.

After 2 years the amount can be found by repeating the interest compounding to obtain

$$b = a\left(1+\frac{r}{2}\right)^4. \tag{12.5}$$

This argument can be continued to obtain the formula for compounding m times per year for t years. If the interest rate is r and it is compounded m times per year for n years then the final amount is

$$b = a\left(1+\frac{r}{m}\right)^{mt}. \tag{12.6}$$

Example 115 *In five years time* \$150 *invested at an interest rate of* 8% *with interest compounded* 4 *times per year and the interest re-invested grows to* $100\left(1+\frac{0.08}{4}\right)^{4\times 5} = \$148.59$.

Equally, if interest is compounded m times per year, an amount b that will be received in n years is today worth

$$a = \frac{b}{\left(1+\frac{r}{m}\right)^{mt}}. \tag{12.7}$$

Now add a table to show how the compounding frequency increases final value. This is the value of \$100 dollars at a annual rate of interest of 10% with different compounding frequencies.

Frequency	Value of \$100
Annually ($m = 1$)	110
Semi-annually ($m = 2$)	110.25
Quarterly ($m = 4$)	110.38
Monthly ($m = 12$)	110.47
Weekly ($m = 52$)	110.61
Daily ($m = 365$)	110.52

Table 12.1: Effect of compounding frequency

The next point is what this table of numbers is converging to. This is continuous compounding which is discussed in the next section.

12.2.3 Continuous Interest

Will make much use of continuous compounding below so introdue it here as an extension of this analysis.

In the limit as we compound more and more frequently we obtain a continuously compounded interest rate. This is because of the formula that

$$\lim_{m\to\infty} a\left(1+\frac{r}{m}\right)^{mt} = ae^{rt}. \tag{12.8}$$

This is continuous compounding. In this expression e is a constant. The value of e is approximately $e = 2.7183$.

Example 116 \$100 *is invested at an interest rate of* 8% *with continuous compounding. The investment grows to* $\$100e^{0.08\times1} = \$108.33$ *after* 1 *year, and* $\$100e^{0.08\times2} = \$117.35$ *after two years.*

It is possible to define equivalent continous and discrete interest rates. This can be done from the following results. Let r_c be the rate of interest with continuous compounding and r_m the rate of interest with discrete compounding m times per year. Then the continuous equivalent to the discrete interest rate is given by

$$r_c = m\ln\left(1+\frac{r_m}{m}\right). \tag{12.9}$$

The discrete equivalent to the continuous interest rate is given by

$$r_m = m\left(e^{r_c/m}-1\right). \tag{12.10}$$

Example 117 *Consider an annual interest rate of* 10% *compounded* 4 *times per year. The continous equivalent is*

$$r_c = 4\ln\left(1+\frac{0.1}{4}\right) = 0.09877.$$

Consider an annual interest rate of 8% *compounded continuously. The equivalent annual interest rate with semi-annual compounding is*

$$r_m = 2\left(e^{0.08/2}-1\right) = 0.081622.$$

12.3 Bonds

A bond is a promise to make certain payments. All bonds are issued with a *maturity date* which is the date at which the final payment is received. (There are some exceptions to this: UK consuls issued to finance Napoleonic War are undated) On the maturity date the bond repays the principal. The principal is also called the *face value*.

As well as the payment of principal, bonds can also make periodic coupon payments. Coupons are typically made semi-annually or annually. The final payment on a bond at the maturity date is the sum of the last coupon and the principal.

12.3.1 Types

There are two distinct categories of bond which differ in whether they make coupon payments or not.

(a) Pure discount bonds.

These are bonds which provide one final payment equal to the face value (or *par value*) of the bond. The return on the bond arises from the fact that they typically sell for less than the face value or "at a discount".

These are the simplest kind of bond and there analysis underlies all other bonds. As noted in the discussion of the efficient frontier, a pure discount bond is basically a simple loan from the bond purchaser to the bond seller with the length of the loan equal to the maturity of the bond. For example, a one-year bond is a one-year loan. This interpretation will be employed frequently in this chapter.

(b)Coupon bonds.

A coupon bond provides a series of payments throughout the life of the bond. These payments are the *coupons* on the bond. It is possible to regard the coupon as an interim interest payment on a loan. This perspective will be found helpful at numerous points below.

So, with a pure discount bond only the final repayment of the loan is made. With a coupon bond, regular interest payments are made then the principal is repaid.

A bond is *callable* if the final payment may be made earlier than maturity. This may sometimes be at a premium meaning the issuer of the bond has to make an additional payment to the holder in order to call. The bond will be called if its issuer finds it advantageous to do so. If it is advantageous for the issuer, it is usually not so for the holder. Hence callable bonds must offer a better return than non-callable to compensate for the risk of calling.

A bond is *convertible* if it includes an option to convert it to different assets. A *sinking fund* is a bond issue which requires that a fraction of the bonds are redeemed each period. This has the advantage of avoiding the necessity for a large payment on the maturity date.

If a Treasury note or Bond is non-callable, it is effectively a portfolio of pure discount bonds. For example, if the bond has a maturity of two, can regard the coupon payment in year 1 as a pure discount and the coupon plus principal in year 2 as a second pure discount. *Coupon stripping* is the process of selling each coupon as an individual asset. This can have the advantage of allowing investors to purchase assets whose timing of payments best matches their needs. Because of this, stripping can create additional value.

12.3.2 Ratings and default

The first way of comparing bonds is to look at ratings. Bonds have some chance of default. This varies across bonds. Government bonds tend to be the safest, while some corporate bonds can be very risky. There are agencies who produce ratings of the riskiness of bonds.

Bonds are rated according to the likelihood of default.

The two most famous rating agencies are:

(1)Standard and Poor's;

(2)Moody's.

The ratings agencies provide use their expertise to provide ratings of debt.

Example 118 *Standard & Poor's is a leading provider of financial market intelligence. The world's foremost source of credit ratings, indices, investment research, risk evaluation and data, Standard & Poor's provides financial decision-makers with the intelligence they need to feel confident about their decisions. Many investors know Standard & Poor's for its respected role as an independent provider of credit ratings and as the home of the S&P 500 benchmark index. But Standard & Poor's global organization also: Provides a wide array of financial data and information, Is the largest source of independent equity research and a leader in mutual fund information and analysis. www.standardandpoors.com*

Example 119 *Moody's Investors Service is among the world's most respected and widely utilized sources for credit ratings, research and risk analysis. Moody's commitment and expertise contribute to stable, transparent and integrated financial markets, protecting the integrity of credit. In addition to our core ratings business, Moody's provides research data and analytic tools for assessing credit risk, and publishes market-leading credit opinions, deal research and commentary, serving more than 9,300 customer accounts at some 2,400 institutions around the globe. www.moodys.com*

The categories used in these ratings systems are:

(1) Moody's assigns bond credit ratings of Aaa, Aa, A, Baa, Ba, B, Caa, Ca, C

(2) Standard & Poor's assign bond credit ratings of AAA, AA, A, BBB, BB, B, CCC, CC, C, D

The top ratins including As are investment grade. Those with Bs are speculative grades. Informally, the lowest category of bonds are known as "junk bonds". These have a very high probability of default.

For corporate bonds, better ratings are associated with:

- Lower financial leverage:
- Smaller intertemporal variation in earnings:
- Larger asset base;
- Profitability;
- Lack of subordination.

The possibility of default implies that a premium must be offered above the risk-free rate of return in order to encourage investors to hold the bonds. This premium is known as the *risk premium.*

12.3.3 Cash and Quoted Prices

The price quoted for a bond is not the price that is paid. The cash price that is paid includes the interest that is accumulated on the bond up to the day of purchase. The quoted price is often called the "clean" price and the cash price is the "dirty" price.

The accumulated interest is calculated on the basis of how many days have passed since the last coupon payment relative to the number of day between coupon payments. There have emerged several alternative ways for doing this count. These are called "Day count conventions". For each bond there is a day count convention that applies.

The accrued interest is defined by the rule

$$\left(\frac{\text{No. of days between dates}}{\text{No. of days in reference period}}\right) \times \text{Interest earned in reference period.} \tag{12.11}$$

The No. of days between dates counts the days from the last coupon payment until the date of sale of the bond. The No. of days in reference period counts the days between the previous coupon payment and the next coupon payment. Interest earned in reference period is the coupon payment.

For Treasury Bonds the day count convention is to use

$$\frac{\text{Actual}}{\text{Actual}} \tag{12.12}$$

Example 120 *Consider a bond with principal* \$100 *that makes coupon payments on March* 1 *and September* 1*. Assume the coupon rate is* 6%*. The accrued interest between March* 1 *and July* 5 *is found from the calculation: Reference period = March* 1 *to September* 1 = 184 *days, interest of* \$3, 126 *days from March* 1 *to July* 5*. So the accrued interest is*

$$\frac{126}{184} \times 3 = 2.0543.$$

For Corporate Bonds the day count convention is

$$\frac{30}{360}. \tag{12.13}$$

What this means is the day count convention assumes there are 30 days per month and 360 days per year.

Example 121 *Consider a bond with principal* $100 *that makes coupon payments on March* 1 *and September* 1*. Assume the coupon rate is* 6%*. The accrued interest between March* 1 *and July* 5 *is found from the calculation: March* 1 *to September* 1 = 180 *days, The number of days between March* 1 *and July* $5 = 4 \times 30 + 4 = 124$*. The accrued interest is*

$$\frac{124}{180} \times 3 = 2.0667.$$

For Money Market Instruments the day count convention is

$$\frac{\text{Actual}}{360}. \tag{12.14}$$

This convention implies that the interest for the whole year is

$$\frac{365}{360} \times \text{interest rate}. \tag{12.15}$$

Example 122 *Consider a bond with principal* $100 *that makes coupon payments on March* 1 *and September* 1*. Assume the coupon rate is* 6%*. The accrued interest between March* 1 *and July* 5 *is found from the calculation: Reference period = March* 1 *to September* 1 = 184 *days, interest of* $3, 126 *days from March* 1 *to July* 5*. So the accrued interest is*

$$\frac{126}{360} \times 6 = 2.1.$$

The price of Treasury bonds is quoted in the US as $ plus thirty-seconds of dollars. Hence the price quote of 90-15 represents 90 15/32 = 90.46875. The price is quoted for face value of $100, so must be scaled up to give the price of a bond with a larger denomination. For example, if the bond has a principal of $100,000 the quoted price of 90-15 becomes $\$90.46875 \times 1000 = \$90,468.75$.

The cash price must include the accrued interest that is calculated using the day count convention. The count convention is designed to give a fair division of the interest between the seller and the purchaser.

In addition, adding the accrued interest ensures that the quoted price does not have any discontinuous jumps. Think of a stock. When it becomes ex dividend the price falls by an amount equal to the dividend. The same would happen to a bond if the quoted price included interest: immediately after the coupon was paid the quoted price would fall by an amount equal to the coupon.

Removing the interest component from the quoted price removes these discontinuous changes. The method of calculation implies that it is the cah price that changes discontinously after coupon payments.

The cash price is determined according to the rule

$$\text{Cash price} = \text{Quoted price} + \text{Accrued interest.} \tag{12.16}$$

Example 123 *Assume it is March* 2, 2007*. Bond is a* 9% *coupon Treasury Bond maturing on July* 12, 2009*. Quoted price is* 94 − 24 (= \$94.75)*. The most recent coupon is January* 12, 2005*, next is on July* 12, 2005 *(*49 *days from January* 12 *to March* 2, 181 *days between coupons). Coupon payment is* \$4.50*. Using the day count convention actual/actual for the Treasury Bond*

$$\frac{49}{181} \times 4.50 = \$1.2182.$$

The cash price per \$100 *is*

$$95.968 = 94.75 + 1.2182.$$

Treasury Bills are sold at a discount so the pricing convention is different. The day count convention is used for treasury bills is

$$\frac{\text{Actual}}{360}. \tag{12.17}$$

If Y is the cash price of a Treasury bill that has n days to maturity the quoted price is

$$\frac{360}{n}(100 - Y). \tag{12.18}$$

The resulting value here is the discount rate: the annualized dollar return provided by the bill expressed as a percentage of face value.

Example 124 *For a* 91*-day bill with cash price* $Y = 98$ *the discount rate is*

$$\frac{360}{91} \times (100 - 98) = 7.91.$$

12.4 Yield-to-Maturity

Bonds available in a large range of maturities and with different strutures of coupon payments. These features make it generally impossible to directly compare the returns offered by different bonds. What is needed to permit comparison is some single number that summarizes the return offered by different bonds.

One measure of return is the promised *yield-to-maturity*. The word "promised" is important here since the bond may be called or go into default. In either case, the full set of promised payments will not be made. The yield-to-maturity is calculated on the basis that neither of these events will occur.

The yield-to-maturity is the most common measure of a bond's return and allows for comparisons between bonds with different structures of payments. The general definition is given by comparing the payments offered by a bond with the payments on an alternative investment at a fixed rate of interest.

Definition 1 *The yield to maturity is the interest rate (with interest compounded at a specified interval) that if paid on the amount invested would allow the investor to receive all the payments of the security.*

The specified interval in the definition corresponds to the frequency of coupon payments made by the bond. The alternative investment can be viewed as placing funds equal to the market price of the bond into a bank account. Interest is paid on this bank account, and withdrawls are made that match the coupon payments of the bond. This definition is now applied to a series of increasingly complex bonds leading to a final general expression.

The notation used is as follows. The principal, or face value, of the bond is denoted M. The principal is paid at the maturity date T. The coupon payment at time t is denoted by C_t and the purchase price of the bond by p. The yield-to-maturity (or just the *yield* from this point onward) is denoted by y.

12.4.1 Discount Bonds

Consider a pure discount bond with principal M and maturity of 1 year. The yield is determined by considering the payments to an investor from the bond and from the alternative investment. The bond can be purchased for price p and one year later the principal M is received. Alternatively, the an investment of p can be made at a fixed rate of interest, y. At the end of the year the investment will be worth $p(1+y)$.

The value of y that ensures these two choices lead to the same final wealth is the yield on the bond. Therefore the yield, y, satisfies the identity

$$p(1+y) = M. \tag{12.19}$$

Example 125 *A bond matures in 1 year, with principal of* \$1000*. If the present price* \$934.58*, the yield-to-maturity satisfies*

$$934.58(1+y) = 1000,$$

so $y = 0.07$ *(7%).*

Next, consider a pure discount bond with a two year maturity. The choices confronting the investor are again to either purchase the bond or invest at a fixed rate of interest, y. Following the latter course of action, the investor will receive interest at the end of the first year to give them a total investment of $p(1+y)$. Re-investing this sum for a second year, interest will again be earned at the end of the second year. The yield then has to satisfy

$$(p(1+y))(1+y) = p(1+y)^2 = M. \tag{12.20}$$

Example 126 *A pure discount bond matures in* 2 *years, with principal* \$1000*. If the present price is* \$857.34*, the yield-to-maturity satisfies*

$$875.34\,(1+y)^2 = 1000,$$

so $y = 0.08$ *(*8%*).*

The expressions can be generalized to conclude that the yield on a pure discount bond with market price p and maturity of T years is the value of y that satisfies

$$p\,(1+y)^T = M. \tag{12.21}$$

12.4.2 Annual Coupons

An investor who purchases a coupon bond receives a flow of coupon payments from the bond until it matures. These coupon payments must be incorporated within the construction of the alternative investment.

Consider a coupon bond with maturity of two years that pays an anuual coupon. Let the principal be M, the coupon payment be C, and the purchase price of the bond be p. The way that the payments on this bond are matched is as follows. The amount p is invested at interest rate y. At the end of the first year after the payment of interest this has become $p\,(1+y)$. The payment of the coupon is equivalent to withdrawing C from this sum. The amount re-invested for a second year is $p\,(1+y) - C$. Interest at rate y is earned on this sum at the end of the second year.

The yield on the bond, y, must then satisfy the identity

$$(p\,(1+y) - C)\,(1+y) = p\,(1+y)^2 - C\,(1+y) = M + C. \tag{12.22}$$

Example 127 *A coupon bond with principal of* \$1000 *pays a coupon of* \$50 *each year and matures in* 2 *years. If the present price is* \$946.93*, the yield-to-maturity satisfies*

$$((1+y)\,946.50 - 50)\,(1+y) = 1050,$$

so $y = 0.08$ *(*8%*).*

Example 128 *A coupon bond with principal of \$100 pays an annual coupon of 10% and matures in 2 years. If the present price is* \$107.33*, the yield-to-maturity satisfies*

$$((1+y)\,107.33 - 10)\,(1+y) = 110,$$

so $y = 0.06$ *(*6%*).*

Reviewing these formula, it can be seen that the last two are particular cases of the expression

$$p = \frac{C}{[1+y]} + \frac{C}{[1+y]^2} + \frac{M}{[1+y]^2}, \tag{12.23}$$

where for the pure discount bond $C = 0$. Observing this, for a bond with maturity of T that pays an annual coupon the general expression defining the yield is

$$p = \sum_{t=1}^{T} \frac{C}{[1+y]^t} + \frac{M}{[1+y]^T}. \tag{12.24}$$

Example 129 *A bond has a maturity of* 10 *years, pays a annual coupon of* \$30 *and has a face value of* \$1000*. If the market price is* \$845.57*, the yield satisfies*

$$845.57 = \sum_{t=1}^{10} \frac{30}{[1+y]^t} + \frac{1000}{[1+y]^{10}},$$

so $y = 0.05$ *(5%).*

It is helpful to add a short explanation of how the yield, y, is actually calculated for these more complex examples. Mathematically, there is no explicit formula that describes the solution for y when $T > 3$. The basic, but time-consuming, approach is to use trial and error. An initial guess of either 5% or 10% is usually worth trying. A more sophisticated approach is to employ a suitable package to graph the value of $p - \sum_{t=1}^{T} \frac{C}{[1+y]^t} - \frac{M}{[1+y]^T}$ as a function of y. The value that makes it equal to zero is the yield, y.

Example 130 *Consider a bond with principal of* \$1000 *that pays an annual coupon of* \$30*. The bond has a maturity of* 5 *years and the current price is* \$800*.*

Using trial and error, produces the following table

y	*0.05*	*0.06*	*0.07*	*0.08*	*0.081*	*0.0801*
$\sum_{t=1}^{5} \frac{30}{[1+y]^t} + \frac{1000}{[1+y]^5}$	913.41	*873.63*	*835.99*	*800.36*	*796.91*	*800.02*

From the table the yield is seen to be $y = 0.0801$.

A graph of $800 - \sum_{t=1}^{5} \frac{30}{[1+y]^t} - \frac{1000}{[1+y]^5}$ *is given in Figure 12.1. This demonstrates the same solution.*

12.4.3 Semi-Annual and More Frequent Coupons

US government bons, and many other government and corporate bonds, make coupon payments on a semi-annual basis. This section shows how these bonds, and indeed bonds that pay coupons at any regular interval, can be incorporated in the framework above.

Consider a bond that pays a semi-annual coupon. Assume each coupon has value C and that the principal is M. Assume that the maturity of the bond is one year. We know from (12.7) that the present value of a payment a made at tn years into the future with an annual interest rate of r compunded m time per year is $a = \frac{b}{\left(1+\frac{r}{m}\right)^{mt}}$. The same rule can be used to discount the payments made by the bond.

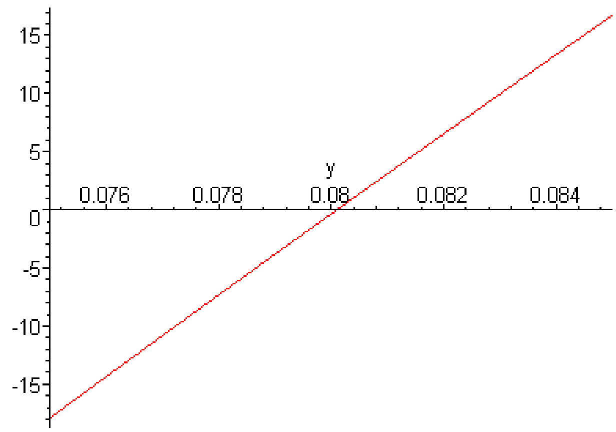


Figure 12.1: Finding the Yield

Semi-annual compounding implies $m = 2$. The first coupon payment is made at time $t = 0.5$. The second coupon and the principal are paid at time $t = 1$. If the price of the bond is p the yield, y, must satisfy

$$p = \frac{C}{\left(1 + \frac{y}{2}\right)^{2\times 0.5}} + \frac{C + M}{\left(1 + \frac{y}{2}\right)^{2\times 1}}.$$

Example 131 *A bond with a maturity of one year pays a coupon semi-annually. Each coupon payment is for* \$3 *and the bond has a maturity value of* \$100*. If its price is* \$92.90*, the yield is defined by*

$$92.90 = \frac{3}{\left(1 + \frac{y}{2}\right)} + \frac{103}{\left[1 + \frac{y}{2}\right]^2},$$

so $y = 0.138\,45$ *(*13.8%*).*

The argument can be extended to bonds with a longer maturity. Consider a bond that pays a semi-annual coupon with maturity of T years. If each coupon payment has value C the yield of the bond satisfies

$$p = \sum_{i=1}^{2T} \frac{C}{\left(1 + \frac{y}{2}\right)^i} + \frac{M}{\left(1 + \frac{y}{2}\right)^{2T}}.$$

Example 132 *A bond with a maturity of* 10 *year pays a* 12% *coupon semi-annually and has a maturity value of* \$100*. The current price of the bond is* \$89.25*. Each coupon payment is for* \$6 *so the yield on the bond satisfies*

$$89.25 = \sum_{i=1}^{20} \frac{6}{\left(1 + \frac{y}{2}\right)^i} + \frac{100}{\left(1 + \frac{y}{2}\right)^{20}},$$

which gives $y = 0.14032$ *(*14%*).*

The analysis can now be extended to more frequent compouding intervals. If the bond pays a coupon m times a year the yield, y, must satisfy

$$p = \sum_{i=1}^{mT} \frac{C}{\left(1+\frac{y}{m}\right)^i} + \frac{M}{\left(1+\frac{y}{m}\right)^{mT}}. \tag{12.25}$$

12.4.4 Continuous Compounding

The yield can also be derived when interest is continously compounded.

Assume the bond pays a coupon C at times $t_1, t_2, ..., t_n$, and that the principal of M is paid at t_n. If interest is continuously compounded and the price of the bond is p the yield satisfies

$$p = Ce^{-yt_1} + Ce^{-yt_2} + ... + Ce^{-yt_n} + Me^{-yt_n}. \tag{12.26}$$

Example 133 *A bond with principal* \$100 *makes a coupon payment of* \$2 *every* 3 *months. The bond has a maturity of* 1 *year and* 3 *months, and a market price of* \$102*. The yield on the bond is given by the solution to*

$$102 = 2e^{-y\times0.25} + 2e^{-y\times0.5} + 2e^{-y\times0.75} + 2e^{-y\times1} + 2e^{-y\times1.25} + 100e^{-y\times1.25}.$$

Hence the yield is $y = 0.062739$ *(6.3%).*

12.4.5 Factors

The yield-to-maturity can be used to determine whether a bond is good value. This can be done by comparing the yield-to-maturity with an estimated appropriate return. In this approach, the investor determines what they feel should be the yield a bond offers and then compares it to the actual yield.

The estimation of an appropriate yield should be based on factors related to the structure of payments and the riskiness of the bonds. The following will be relevant:

- *Time to maturity* This will determine the time until the principal is received. If the case is required earlier, the bond will have to be sold.
- *Coupon payments* Coupon payments relate to the timing of the payment flow compared to preferred flow
- *Call provisions* To see the effect of these, consider why a bond would be called. As an example, assume that on issue the bond had a coupon of \$100 but now similar bonds have coupon of \$50. It would pay the issuer to call the bond and replace it with the lower coupon bond. So, bonds are generally called when yields fall. This benefits the issuer but not the purchaser. Hence bonds with call provision should have higher yields to compensate.
- *Tax status* Bonds which are tax exempt will have a lower yield-to-maturity that reflects the tax advantage.

- *Marketability (Liquidity)* Bonds that are not very marketable need to have a higher yield-to-maturity to induce investors to purchase them.

- *Likelihood of default* As already described, bonds that may default have a risk premium so the yield-to-maturity is higher.

Taken together, these factors determine an overall view of the bond and from this the appropriate return can be inferred.

12.5 Bond Properties

This is a collection of stuff on bonds.

12.5.1 Duration

Duration is a measure of the length of time until the average payment is made on a bond. Its can be used to compare different bonds. Duration can also be used to capture the sensitivity of price to the interest rate. This section shows hwo to calculate duration for a single bond and then for a portfolio of bonds.

If cash flows are received at times $1, ..., T$ then the duration, D, is given by

$$D = \frac{PV(1) + 2 \times PV(2) + ... + T \times PV(T)}{PV} \tag{12.27}$$

where $PV(t)$ is the present value of the cash flow at time t and is defined by

$$PV(t) = \frac{C_t}{(1+y)^t}, \tag{12.28}$$

and PV is the total present value of the cash flow. When this formula is applied to a bond, the pricing ensures that PV is also the market price of the bond.

For a zero-coupon bond no payments are made prior to the final value. Hence $PV(T) = PV$ so

$$D = T, \tag{12.29}$$

and the duration is equal to the time to maturity. For a coupon bond, the intermediate payments ensure that the duration has to be less than the maturity, giving

$$D < T. \tag{12.30}$$

Example 134 *Consider a bond that pays an annual coupon of \$40, has a face value of \$1000 and a maturity of 6 years. With a discount rate of 3%, the following table computes the values required to calculate the duration*

Time	*Cash Flow*	*Discount Factor*	*PV of Cash Flow*	$\times t$
1	*40*	$\frac{1}{1.03} = 0.97087$	$40 * 0.97087 = 38.835$	38.835
2	*40*	$\frac{1}{[1.03]^2} = 0.94260$	$40 * 0.94260 = 37.704$	$2 * 37.704 = 75.408$
3	*40*	$\frac{1}{[1.03]^3} = 0.91514$	$40 * 0.91514 = 36.606$	$3 * 36.606 = 109.82$
4	*40*	$\frac{1}{[1.03]^4} = 0.88849$	$40 * 0.88849 = 35.540$	$4 * 35.540 = 142.16$
5	*40*	$\frac{1}{[1.03]^5} = 0.86261$	$40 * 0.86261 = 34.504$	$5 * 34.504 = 172.52$
6	*1040*	$\frac{1}{[1.03]^6} = 0.83748$	$1040 * 0.83748 = 870.98$	$6 * 870.98 = 5225.9$

Using these values the duration is

$$\frac{\sum PV * t}{\sum PV} = \frac{38.835 + 75.408 + 109.82 + 142.16 + 172.52 + 5225.9}{38.835 + 37.704 + 36.606 + 35.540 + 34.504 + 870.98} = 5.4684.$$

The calculation of the duration can be extended to portfolios of bonds. Consider two bonds A and B with durations

$$D^A = \frac{\sum_{t=1}^{T} tPV^A(t)}{PV^A}, \tag{12.31}$$

where $PV^A = \sum_{t=1}^{T} PV^A(t)$, and

$$D^B = \frac{\sum_{t=1}^{T} tPV^B(t)}{PV^B}, \tag{12.32}$$

with $PV^B = \sum_{t=1}^{T} PV^B(t)$.

These facts imply that

$$PV^A D^A + PV^B D^B = \sum_{t=1}^{T} tPV^A(t) + \sum_{t=1}^{T} tPV^B(t). \tag{12.33}$$

The duration of the portfolio is defined by

$$D = \frac{\sum_{t=1}^{T} tPV^A(t) + \sum_{t=1}^{T} tPV^B(t)}{PV}, \tag{12.34}$$

where $PV = PV^A + PV^B$. But (12.33) implies that

$$D = \frac{PV^A}{PV} D^A + \frac{PV^B}{PV} D^B. \tag{12.35}$$

This result establishes that the duration of a portfolio is a weighted sum of durations of the individual bonds.

12.5.2 Price/Yield Relationship

From the fact that the price of the bond is determined by

$$P = \frac{C}{1+y} + \frac{C}{(1+y)^2} + ... + \frac{C+M}{(1+y)^T}, \tag{12.36}$$

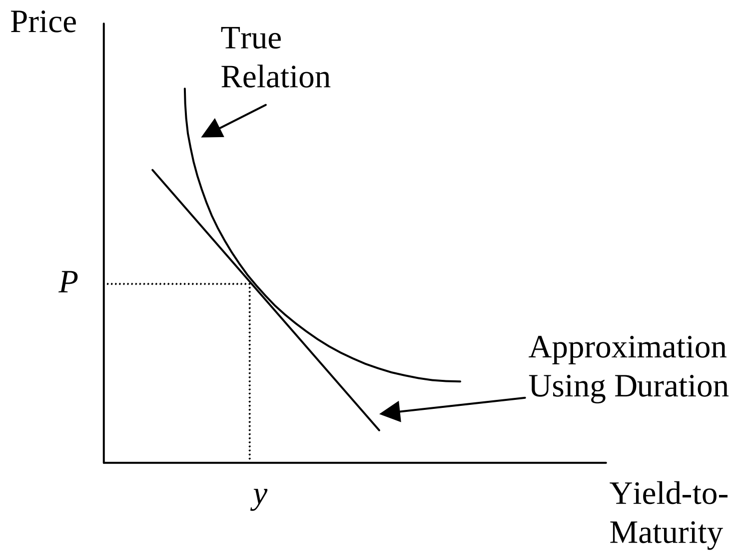


Figure 12.2: Price/Yield Relationship

it can be observed that:

1. P and y are inversely related.
This follows from seeing that

$$\frac{dP}{dy} = -\frac{C}{(1+y)^2} - \frac{2C}{(1+y)^3} - ... - \frac{T(C+M)}{(1+y)^{T+1}} < 0. \quad (12.37)$$

2. The relationship is convex.
Calculation gives

$$\frac{d^2P}{dy^2} = \frac{2C}{(1+y)^3} + \frac{3C}{(1+y)^4} + ... + \frac{(T+1)(C+M)}{(1+y)^{T+2}} > 0. \quad (12.38)$$

The duration can also be used to link changes in the yield on a bond to changes in its price. From (12.28)

$$\frac{dPV(t)}{dy} = -\frac{t}{1+y}PV(t). \quad (12.39)$$

Hence using the fact that $PV = P \equiv \sum_{t=1}^{T} PV(t)$,

$$\frac{dP}{dy} = \frac{d\sum_{t=1}^{T} PV(t)}{dy} = -\sum_{t=1}^{T} \frac{t}{1+y} PV(t) = -\frac{1}{1+y} DP, \quad (12.40)$$

or

$$\frac{dP}{dy} = -D_m P, \quad D_m = \frac{1}{1+y} D, \quad (12.41)$$

where D_m is called the *modified duration*.

This is an exact result that holds for a differential (meaning very small change) in the yield. If can also be used as an approximation relating the price to the modified duration and yield changes. So

$$\Delta P \approx -D_m P \Delta y. \tag{12.42}$$

This shows the approximation, but comparison with (12.38) also shows that the use of duration overstates the effects of a yield increase.

This result is returned to later after the yield curve has been considered.

12.6 Bond Portfolios

Some stuff here on the management of bond protfolis.

12.6.1 Immunisation

Duration of a portfolio

12.6.2 Hedging

Using bonds to hedge

12.7 Conclusions

The chapter has considered methods for comparing bonds with different structures of payments and different maturities. Bond ratings were analyzed as was the yield as a measure of the return. Bonds represent one form of lending, so the interest rates on bonds are related to the interest rates on loans. This analysis tied together spot rates, forward rates and discount factors. The duration as another measure of a bond was also considered and price/yield relationships were investigated.

Exercise 88 *What is "coupon stripping"? What are the benefits of this for investors?*

Exercise 89 *Three pure discount bonds, all with face values of $1000, and maturities of 1, 2 and 3 years are priced at $940, $920 and $910 respectively. Calculate their yields. What are their yields if they are coupon bonds with an annual coupon of $40?*

Exercise 90 *An investor looks for a yield to maturity of 8% on fixed income securities. What is the maximum price the investor would offer for a coupon bond with a $1000 face value maturing in 3 years paying a coupon of $10 annually with the first payment due one year from now? What is the maximum price if it is a pure discount bond?*

Exercise 91 *Three pure discount bonds, all with face values of $1000, and maturities of 1, 2 and 3 years are priced at $950.89, $942.79 and $929.54 respectively. Calculate:*

a. The 1-year, 2-year and 3-year spot rates;

b. The forward rates from year 1 to year 2 and from year 2 to year 3.

Exercise 92 *Calculate the duration of a bond with a coupon of $50 and maturity value of $1000 if it matures in six years and the discount rate is 4%.*

Chapter 13

The Term Structure

13.1 Introduction

This chapter looks at the variation of yields with respect to time and reviews theories designed to explain this.

13.2 Yield and Time

The *yield curve* shows the yield-to-maturity for treasury securities of various maturities at a particular date. In practice, securities do not lie exactly on this line because of differences in tax treatment and in callability.

It should be noted that the yield-to-maturity is a derived concept from the flow of payments and it would equally informative to have used duration on horizontal axis rather than maturity.

13.3 Interest Rates and Discounting

There are a series of interest rates in the market place. These must be related to prevent arbitrage. Such arbitrage would involve constructing an arbitrage portfolio of loans. This section now relates these. It also ties in with the idea of discounting.

13.3.1 Spot Rates

The *spot rate* is the interest rate associated with a spot loan: a loan that is granted immediately ("on the spot") with capital and interest repaid at a specified date. The discussion of the efficient frontier in Chapter 4 has already made the interpretation of a bond as a loan. So the spot rates must be related to the yields on bonds.

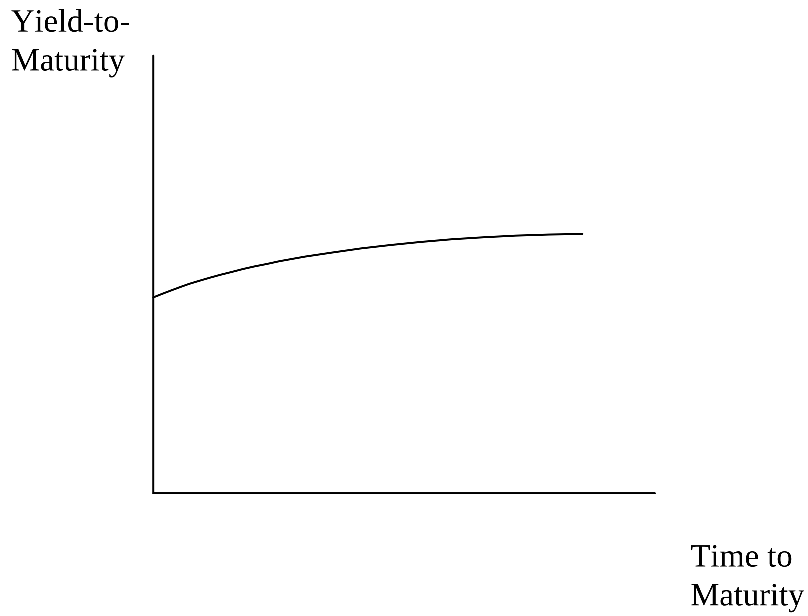


Figure 13.1: Yield Curve

For pure discount bonds the relationship is very straightforward. A pure discount bond is simply a loan from the purchaser to the issuer with the length of the loan equal to the maturity of the bond. The yield on the bond must therefore be equal to the rate of interest on a spot loan of this length. This gives the identity

$$\text{spot rate} = \text{yield-to-maturity}. \tag{13.1}$$

This identity is true of any pure discount bond. Therefore the price of a discount bond with maturity T is related to the spot rate S_t by

$$p = \frac{M}{(1 + S_T)^T}. \tag{13.2}$$

Given a set of pure discount bonds of maturities $T = 1, 2, ...$ the application of this formula provides the spot rates $S_1, S_2, ...$

Example 135 *Three pure discount bonds of with principal of \$1000 and with maturities of 1, 2 and 3 years have prices \$934.58, \$857.34, \$772.18 respectively. The corresponding spot rates found from* $934.58 = \frac{1000}{1+S_1}, 857.34 = \frac{1000}{(1+S_2)^2}$ *and* $772.18 = \frac{1000}{(1+S_3)^3}$*. Hence* $S_1 = 0.07$ *(7%),* $S_2 = 0.08$ *(8%) and* $S_3 = 0.09$ *(9%). Therefore an interest rate of 7% is paid on an immediate loan to be re-paid in 1 year and an interest rate of 9% applies to an immediate loan which has to be re-paid in 3 years.*

The analysis has to be adjusted to apply to a coupon bond. In the discussion of coupon stripping it was noted how each coupon payment could be treated as a separate loan. Hence the coupon paid at the end of the first year can be treated as repayment of principal on a 1-year loan. This should attract interest at rate S_1. Similarly, the coupon payment at the end of the second year can be treated as the repayment of principal on a 2-year loan. It attracts interest at rate S_2. The same logic can be applied to all later payments. The price paid for the bond must represent the sum of the values of the repayments on these individual loans. Hence the relationship of the price, coupon payments and principal to the spot rates for a coupon bond of maturity T is given by

$$p = \sum_{t=1}^{T} \frac{C}{(1+S_t)^t} + \frac{M}{(1+S_T)^T}. \tag{13.3}$$

Example 136 *A bond with maturity of 3 years has a principal of \$1000 and makes a coupon payment of \$50. If the price is \$900 then the spot rates satisfy*

$$900 = \frac{50}{1+S_1} + \frac{50}{[1+S_2]^2} + \frac{50}{[1+S_3]^3} + \frac{1000}{[1+S_3]^3}.$$

It is clear that the spot rates cannot be calculated using information on a single coupon bond. Instead they must be constructed by an iterative process. This works by first taking either a pure discount bond or a coupon bond with a maturity of 1 year. With the pure discount bond, S_1 is determined by $p = \frac{M}{1+S_1}$ and with the coupon bond by $p = \frac{C}{1+S_1} + \frac{M}{1+S_1}$. The spot rate S_2 can then be found using coupon bond with a maturity of 2 years by observing that $p = \frac{C}{1+S_1} + \frac{C}{[1+S_2]^2} + \frac{M}{[1+S_2]^2}$ can be solved for S_2 once S_1 is known. Next, using S_1 and S_2 it is possible to use a coupon bond with a maturity of 3 years to find S_3. This process can be continued to consecutively construct a full set of spot rates using the prices of a series of bonds of different maturity.

Example 137 *Three bonds have face values of \$1000. The first is a pure discount bond with price \$909.09, the second is a coupon bond with coupon payment \$40, a maturity of 2 years and a price of \$880.45 and the third bond is a coupon bond with coupon of \$60, maturity of 3 years and price of \$857.73.*

The spot rate S_1 is given by

$$909.09 = \frac{1000}{1+S_1},$$

so $S_1 = 0.1$ (10%). Using the fact that $S_1 = 0.1$, S_2 is determined by

$$880.45 = \frac{40}{1.1} + \frac{1040}{[1+S_2]^2},$$

so $S_2 = 0.11$ (11%). Finally, using S_1 and S_2, S_3 solves

$$857.73 = \frac{60}{1.1} + \frac{60}{[1.11]^2} + \frac{1060}{[1+S_3]^3}.$$

This gives $S_3 = 0.12$ (12%).

13.3.2 Discount Factors

If a future flow of payments are to be received, the standard process is to convert these to a present value by using discounting. The reason for doing this is that it allows different flows to be directly compared by using their present values. Discount factors are used to discount payments to find the present value of future payments.

These discount factors can be used to find the present value of any security. Let the payment in period t be V_t and assume the last payment is received in period T. With discount factor d_t the present value of the flow of payments is

$$PV = \sum_{t=1}^{T} d_t V_t. \tag{13.4}$$

This can also be expressed in terms of a discount rate. If the discount rate, ρ, is constant then $d_t = \frac{1}{(1+\rho)^t}$ and

$$PV = \sum_{t=1}^{T} \frac{1}{(1+\rho)^t} V_t. \tag{13.5}$$

This method of discounting can be used whether the payments are certain or risky. When they are risky it is necessary to take explicit account of the risk. One way to do this was seen in Chapter 8 where the expected value of the payment in period t was used and the discount rate adjusted for risk using the beta. An alternative way of incorporating risk in the discounting will be see in Chapter 14.

If the payments are certain, then the there is no need to adjust for risk and the discount factors can be related directly to the returns on bonds and the spot rates. In fact, if d_t is defined as the present value of \$1 in t years, then

$$d_t = \frac{1}{(1+S_t)^t}, \tag{13.6}$$

where S_t is the spot rate on a loan that must be repaid in t years. If the value of \$1 were above or below this value, then an arbitrage possibility would arise. Using these discount factors, the present value of a flow of payments is

$$PV = \sum_{t=1}^{T} \frac{1}{(1+S_t)^t} V_t. \tag{13.7}$$

Example 138 *If* $d_1 = 0.9346$ *and* $d_2 = 0.8573$ *and a security pays \$70 in 1 years time and \$1070 in 2 years time then*

$$P = 0.9346 \times 70 + 0.8573 \times 1070 = 982.73.$$

Example 139 *If* $S_1 = 0.09, S_2 = 0.1$ *and* $S_3 = 0.11$, *the present value of the flow* $V_1 = 50, S_2 = 50$ *and* $S_3 = 1050$ *is*

$$PV = \frac{50}{1.09} + \frac{50}{(1.1)^2} + \frac{1050}{(1.11)^3} = 854.94.$$

13.3.3 Forward Rates

The spot rates of interest relate to immediate loans. It is also possible to consider agreeing today for a loan to be granted at some future date with repayment at some even later data. For instance, an investor could agree to receive a loan in one year's time to be paid back in two years. Such loans are called *forward loans.*

The rate of interest on a forward loan is called the *forward rate.* The interest rate on the loan made in one year's time to be paid back in two years is denoted by $f_{1,2}$. It should be stressed that this is an interest rate agreed today for a loan in the future. If the loan contract is accepted by the lender and borrower this is the interest rate that will be paid on that loan. The important point is that it need not be the same as the rate of interest that applies to one-year spot loans in a year's time.

Forward rates have to be related to current spot rates to prevent arbitrage, so they link the spot rates for different years. To see how this link emerges, consider two alternative strategies:

- Invest for one year at spot rate S_1 and agree today to invest for a second year at forward rate $f_{1,2}$;
- Invest for two years at spot rate S_2.

To avoid any possibility of arbitrage, the returns on these two strategies must be equal. If they were not, it would be possible to borrow at the lower rate of interest and invest at the higher, yielding a risk-free return for no net investment. A dollar invested in strategy 1 is worth $(1+S_1)$ after one year and, reinvested at interest rate $f_{1,2}$, becomes $(1+S_1)(1+f_{1,2})$ after two years. A dollar invested in strategy 2 is worth $(1+S_2)^2$ after two years. The equality between the returns requires that

$$(1+S_1)(1+f_{1,2}) = (1+S_2)^2. \tag{13.8}$$

Hence

$$1+f_{1,2} = \frac{(1+S_2)^2}{(1+S_1)}. \tag{13.9}$$

The spot rates therefore determine the interest rate on a forward loan.

Example 140 *Let* $S_1 = 0.08$ *and* $S_2 = 0.09$. *Then*

$$1+f_{1,2} = \frac{(1+0.09)^2}{(1+0.08)},$$

so $f_{1,2} = 0.1$.

The same argument can be applied between to link the spot rates in any periods t and $t-1$ to the forward rate $f_{t,t-1}$. Doing so gives the general formula for the forward rate between years $t-1$ and t as

$$1 + f_{t,t-1} = \frac{(1 + S_t)^t}{(1 + S_{t-1})^{t-1}}. \tag{13.10}$$

Example 141 *The spot rate on a loan for 10 years is 12% and the spot rate on a loan for 11 years is 13%. To prevent arbitrage, the forward rate* $f_{10,11}$ *has to satisfy*

$$1 + f_{10,11} = \frac{(1 + 0.13)^{11}}{(1 + 0.12)^{10}},$$

so $f_{10,11} = 23.5\%$.

Forward rates are linked to spot rates and spot rates determine discount factors. Therefore there is a link between forward rates and discount factors. This is given by the relation

$$d_t = \frac{1}{(1 + S_{t-1})^{t-1}(1 + f_{t-1,t})}. \tag{13.11}$$

13.4 ***Converting Interest Rates***

The issue that remains is to convert the semi-annual or monthly interest rates into annual equivalents. There are two ways that this can be done which lead to slightly different answers.

To motivate the first of these, consider investing \$1 for a year with interest paid semi-annually at rate y. After 6 months, the \$1 becomes $\$1 + y$ after the payment of interest. Now assume that the interest is not reinvested, but the capital sum of \$1 is. At the end of the year the \$1 that has been invested has become $\$1 + y$. Adding the interest of $\$y$ that was withdrawn after 6 months, gives the investor at total of $\$1 + 2y$. The annual interest rate can then be interpreted as $2y$. Under this first approach the semi-annual interest rate is converted to an annual rate by multiplying by 2. More generally, if interest is paid n times per year at rate y, the annual interest rate is ny.

The second method of converting to an annual rate is to assume that the interest earned after 6 months is reinvested. After 6 months, the \$1 investment is worth $\$1 + y$, and with reinvestment is worth $\$(1 + y)(1 + y)$ after 1 year. This corresponds to an annual interest rate of $(1 + y)^2 - 1$.

Example 142 *Assume the semi-annual interest rate is 5%. Without reinvestment, the annual interest rate is* $2 \times 5 = 10\%$. *With reinvestment, the annual interest rate is* $(1 + 0.05)^2 - 1 = 0.1025$ *(10.25%).*

Example 143 *An investment pays interest of 2% each month. Without reinvestment, the annual interest rate is 24%. With reinvestment it is* $(1 + 0.02)^{12} - 1 = 0.26824$ *(26.824%).*

As the examples illustrate, the annual interest rate with reinvestment is higher than without. For semi-annual interest, the difference is given by

$$(1+y)^2 - 1 - 2y = y^2. \tag{13.12}$$

When $y = 0.05$, the difference $y^2 = 0.0025$, as found in the example. In general, if interest is paid n times per year, the difference between the annual interest rate with reinvestment and that without reinvestment is

$$(1+y)^n - 1 - ny. \tag{13.13}$$

Example 144 *If interest of 1% is paid monthly, the difference between the two annual interest rates is*

$$(1+0.01)^{12} - 1 - 12 \times 0.01 = 0.006825.$$

There is no right or wrong in which of these interest rates to use. Both are derived from legitimate, though different, experiments. In the range of interest rates usually encountered in practice, the difference is small but significant. When such conversions are necessary in later parts of the text, the reinvestment method will be used for simplicity.

13.5 Term Structure

A similar graph can be constructed using spot rates on the vertical axis. This is called the *term structure* of interest rates. Spot rates are more fundamental than the yield-to-maturity.

The following question are raised by the term structure:

i. Why do rates vary with time?

ii. Should the term structure slope up or down?

Although the term structure can slope either way, periods in which it slope upwards are more common than periods in which it slopes down.

The following theories have been advanced to answer these questions.

13.6 Unbiased Expectations Theory

This theory is based on the view that forward rates represent an average opinion about expected rates in the future. So,

- if yield curve upward sloping, rates are expected to rise,
- if yield curve downward sloping, rates are expected to fall.

Example 145 *Consider the investment of £1. Let the 1-year spot rate be 7%, the two year spot rate be 8%.*

Consider the following two strategies

a. invest now for two years.

Final return $= 1 \times [1.08]^2 = 1.664$

b. invest for one year, then again for a further year
Final return $= 1 \times [1.07]\,[1 + es_{1,2}]$
where $es_{1,2}$ *is the expected one year spot rate in year 2.*
This strategies must yield the same return which implies $es_{1,2} = 0.0901$ *.*
Reversing this argument
- one year spot rate today is 7%
- one year spot rate expected next year is 9.01%
- so two year spot rate must be 9%
Hence the yield curve slopes upwards under the assumptions.

In equilibrium: it must be the case that

$$es_{1,2} = f_{1,2}, \tag{13.14}$$

so that the expected future spot rate is equal to the forward rate. This would be true for all time periods.

13.7 Liquidity Preference Theory

This theory is based on the idea that investors prefer, all things equal, short-term securities to long-term securities. This can be justified by assuming that investors place an intrinsic value on liquidity.

For example, consider making an investment for a two-year period. This can be done using two different strategies.
i. Maturity Strategy
- hold a two-year asset
ii. Rollover Strategy
- hold two one-year assets

An investor who values liquidity would prefer the rollover strategy. They might need cash at end of period 1 and with maturity strategy, price of asset at end of year 1 is not known. Using the rollover strategy eliminates this price risk. Consequently, in order to make them attractive, longer term securities must have a risk premium

To see this
- expected return on £1 with rollover strategy is

$$1 \times [1 + s_1]\,[1 + es_{1,2}] \tag{13.15}$$

- expected return on £1 with maturity strategy is

$$1 \times [1 + s_2]^2 \tag{13.16}$$

The maturity strategy must have higher return to compensate for loss of liquidity so

$$[1 + s_1]\,[1 + es_{1,2}] < [1 + s_2]^2 \tag{13.17}$$

Since by definition

$$[1 + s_1]\,[1 + f_{1,2}] = [1 + s_2]^2 \qquad (13.18)$$

it follows that

$$f_{1,2} > e_{1,2} \qquad (13.19)$$

or

$$f_{1,2} = e_{1,2} + L_{1,2} \qquad (13.20)$$

where $L_{1,2}$ is the liquidity premium.

Under the Liquidity Preference Theory the term structure again depends on the expected spot rate but with the addition of the liquidity premium.

Note that if all spot rates are equal, the liquidity premium ensures the term structure slopes upwards. For it to slope downwards, spot rates must be falling. Thus the liquidity premium ensures that the term structure slopes upwards more often than it slopes downwards.

13.8 Market Segmentation (Preferred Habitat)

The basic hypothesis of this theory is that the market is segmented by maturity date of the assets. It motivates this by assuming that investors have different needs for maturity.

The consequence is that supply and demand for each maturity date are independent and have their return determined primarily by the equilibrium in that section. Points on the term structure are related only by substitution of marginal investors between maturities.

13.9 Empirical Evidence

Strict market segmentation - little empirical support

- observe continuity of term structure

There is some evidence that term structure conveys expectations of future rates

- but with inclusion of liquidity preferences
- but these premiums change over time

13.10 Implications for Bond Management

Look at how bonds can be managed to protect against effects of interest changes.

Link the duration, and interest rates and term structure

Add immunization methods.

13.11 Conclusion

Complete explanation has not been found but term structure can be used to provide information on expected level of future rates.

Exercise 93 *Derive a term structure.*

Exercise 94 *Solve an immunization example.*

Exercise 95 *Do price/yield and duration example.*

Exercise 96 *Example on risk minimization.*

Part VI

Derivatives

Chapter 14

Options

In the practise of investment analysis, as in life generally, commitment can be costly. Commitment can force a damaging course of action to be seen through to the end long after it is clear that it is wrong. Options, though, are valuable. They allow us to pick what is right when it is right or to choose not to select anything at all. Simple though it may be, the act of "keeping our options open" is good investment advice. Financial markets have long realized these facts and have developed financial instruments that allow options to be kept open. Since having an option is valuable, it can command a price and be traded on a market. The purpose of investment analysis is to determine the value to place upon an option. This may seem an imprecise question, but in no other area of finance has investment analysis been more successful in providing both a very clear answer and revolutionizing the functioning of the market.

14.1 Introduction

An option is a contract that gives the holder the right to undertake a transaction if they wish to do so. It also gives them the choice to not undertake the transaction. Possessing this freedom of choice is beneficial to the holder of the option since they can avoid being forced to make an undesirable trade. Options therefore have value and the rights to them are marketable.

The issue that the investment analyst must confront when faced with options is to determine their value. It is not possible to trade successfully without knowing the value of what is being traded. This applies equally to the financial options traded on established markets and to more general instruments, such as employment contracts, which have option-like features built in. This chapter will describe the standard forms of option contract and then gradually build towards a general formula for their valuation. The individual steps of the building process have independent worth since they provide a methodology for tackling a range of valuation issues.

14.2 Options

There are two basic types of options. A *call option* gives the right to buy an asset at a specific price within a specific time period. A *put option* gives the right to sell an asset at a specific price within a specific time period. The price at which the trade can take place is called the *exercise* or *strike* price. The asset for which there is an option to buy or sell is often called the *underlying asset*. If the underlying asset is a common stock, then the standard call and put options are called *plain vanilla options*. This distinguishes them from other more complex options which, for example, can provide the option to buy another option. If the option is used, for example the holder of a call option chooses to buy the underlying stock, the option is said to have been *exercised*.

14.2.1 Call Option

A plain vanilla call option is the right to buy specific shares for a given price within a specified period. The *premium* on an option is the price paid by the investor to purchase the option contract.

The contract for a plain vanilla call option specifies:

- The company whose shares are to be bought;
- The number of shares that can be bought;
- The purchase (or *exercise*) price at which the shares can be bought;
- The date when the right to buy expires (*expiration date*).

A *European call* can only be exercised at the time of expiration. This means that the purchaser of the option must hold it until the expiration date is reached and only then can choose whether or not to exercise the option. In contrast an *American call* can be exercised at any time up to the point of expiration.

If an investor purchases a call option, they must have some expectation that they will wish to exercise the option. Whether they will wish to do so depends critically upon the relationship between the exercise price in the contract and the price of the underlying asset. Clearly they will never exercise the right to buy if the price of the underlying asset is below the exercise price: in such a case they could purchase the underlying asset more cheaply on the standard market.

For a European call, the option will always be exercised if the price of the underlying asset is above the exercise price at the date of expiration. Doing so allows the investor to purchase an asset for less than its trading price and so must be beneficial. With an American call, the issue of exercise is more complex since there is also the question of when to exercise which does not arise with European options. Putting a detailed analysis of this aside until later, it remains correct that an American option will only be exercised if the price of the underlying is above the exercise price and will certainly be exercised if this is true at the expiration date.

Example 146 *On July 11 2003 Walt Disney Co. stock were trading at $20.56. Call options with a strike or exercise price of $22.50 traded with a premium of $0.05. These call options will only be exercised if the price of Walt Disney Co. stock rises above $22.50.*

In order for a profit to be made on the purchase of a call option it is necessary that the underlying stock prices rises sufficiently above the exercise price to offset the premium.

Example 147 *Call options on Boeing stock with a strike price of $30.00 were trading at $5.20 on June 23, 2003. If a contract for 100 stock were purchased this would cost $520. In order to make a profit form this, the price on the exercise date must be above $35.20.*

The next example describes the financial transfers between the two parties to an options contract.

Example 148 *Consider A selling to B the right to buy 100 shares for $40 per share at any time in the next six months. If the price rises above $40, B will exercise the option and obtain assets with a value above $40. For example, if the price goes to $50, B will have assets of $5000 for a cost of $4000. If the price falls below $40, B will not exercise the option. The income for A from this transaction is the premium paid by B to purchase the option. If this is $3 per share, B pays A $300 for the contract. If the price of the share at the exercise date is $50, the profit of B is $5000 - $4000 - $300 = £700 and the loss of A is £300 - $1000 = $700. If the final price $30, the profit of A is $300 and the loss of B is - $300.*

Two things should be noted from this example. Firstly, the profit of one party to the contract is equal to the loss of the other party. Options contracts just result in a direct transfer form one party to the other. Secondly, the loss to A (the party selling the contract) is potentially unlimited. As the price of the underlying stock rises, so does their loss. In principle, there is no limit to how high this may go. Conversely, the maximum profit that A can earn is limited to the size of the premium.

The final example illustrates the general rule that call options with lower exercise prices are always preferable and therefore trade at a higher price. Having a lower exercise price raises the possibility of earning a profit and leads to a greater profit for any given price of the underlying.

Example 149 *On June 23, 2003 IBM stock were trading at $83.18. Call options with expiry after the 18 July and a strike price of $80 traded at $4.70. Those with a strike price of $85 at $1.75.*

A final point needs to be noted. Exercise of the option does not imply that the asset is actually sold by one party to the other. Because of transactions costs, it is better for both parties to just transfer cash equal in value to what would happen if the asset were traded.

14.2.2 Put Options

A plain vanilla put option is the right to sell specific shares for a given price within a specified period. The contract for a plain vanilla put option specifies:

- The company whose shares are to be sold;
- The number of shares that can be sold;
- The selling (or *exercise*) price at which the shares can be sold;
- The date when the right to sell expires (*expiration date*).

As with calls, a *European put* can only be exercised at the expiration date whereas an *American put* can be exercised at any date up to the expiration date. The difference in value between American puts and European puts will be explored later. But it can be noted immediately that since the American put is more flexible than the European put, it value must be at least as high.

Example 150 *On July 11 2003 Walt Disney Co. stock were trading at $20.56. Put options with a strike or exercise price of $17.50 traded with a premium of $0.10. These put options will only be exercised if the price of Walt Disney Co. stock falls below $17.50.*

It is only possible to profit from purchasing a put option if the price of the underlying asset falls far enough below the exercise price to offset the premium.

Example 151 *Put options on Intel stock with a strike price of $25.00 were trading at $4.80 on June 23, 2003. If a contract for 100 stock were purchased this would cost $480. In order to make a profit form this, the price on the exercise date must be below $20.20.*

In contrast to the position with a call option, it can be seen from the next example that the loss to the seller of a put contract is limited, as is the potential profit for the purchaser. In fact, the loss to A (or profit to B) is limited to the exercise price and the loss of B (profit to A) is limited to the premium.

Example 152 *A sells B the right to sell 300 shares for $30 per share at any time in the next six months. If the price falls below $30, B will exercise the option and obtain a payment in excess of the value of the assets. For example, if the price goes to $20, B will receive $9000 for assets worth $6000. If the price stays above $30, B will not exercise the option. The income for A is the premium paid by B for the option. If this is $2 per share, B pays A $600 for the* *contract. If the price of the stock at expiry of the contract is \$20, the profit of B is \$9000 - \$6000 - \$600 = \$2400 and the loss of A is \$6000 + \$600 - \$9000 = - \$2400. If the final price is \$40, the loss of B is - \$600 and the profit of A is \$600.*

The next example illustrates that the higher is the strike price, the more desirable is the put option. This is because a greater profit will be made upon exercise.

Example 153 *On June 23, 2003 General Dynamics stock were trading at \$73.83. Put options with expiry after the 18 July and a strike price of \$70 traded at \$1.05. Those with a strike price of \$75 traded at \$2.95.*

14.2.3 Trading Options

Options are traded on a wide range of exchanges. Most prominent amongst those in the US are the Chicago Board Options Exchange, the Philadelphia Stock Exchange, the American Stock Exchange and the Pacific Stock Exchange. Important exchanges outside the US include the Eurex in Germany and Switzerland and the London International Financial Futures and Options Exchange.

Options contracts are for a fixed number of stock. For example, an options contract in the US is for 100 stock. The exercise or strike prices are set at discrete intervals (a \$2.50 interval for stock with low prices, up to a \$10 interval for stock with high prices). At the introduction of an option two contracts are written, one with an exercise prices above the stock price and one with an exercise price below. If the stock price goes outside this range, new contracts can be introduced. As each contract reaches its date of expiry, new contracts are introduced for trade.

Quotes of trading prices for options contracts can be found in both The Wall Street Journal and the Financial Times. These newspapers provide quotes for the call and put contracts whose exercise prices are just above and just below the closing stock price of the previous day. The price quoted is for a single share, so to find the purchase price of a contract this must be multiplied by the number of shares in each contract. More detailed price information can also be found on Yahoo which lists the prices for a range of exercise values, the volume of trade and the number of open contracts.

Market makers can be found on each exchange to ensure that there is a market for the options. The risk inherent in trading options requires that margin payments must be must in order to trade.

14.3 Valuation at Expiry

The value of an option is related to the value of the underlying asset. This is true throughout the life of an option. What is special about the value of the option at the expiration date is that the value can be computed very directly.

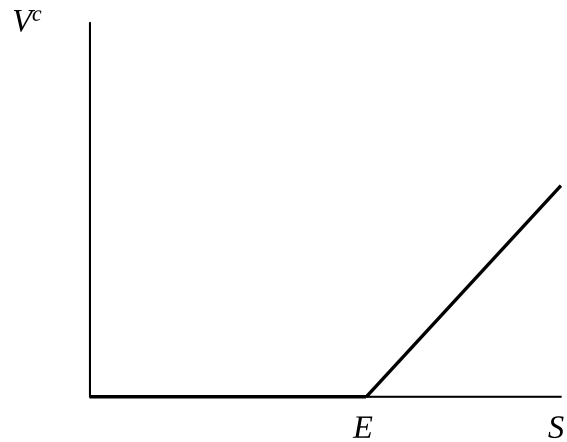


Figure 14.1: Value of a Call Option at Expiry

Prior to expiration the computation of value requires additional analysis to be undertaken, with the value at expiration an essential component of this analysis.

If a call option is exercised the holder receives a sum equal to the difference between the price of the underlying stock at expiry and the exercise price. An option with an exercise price of \$10 on an underlying stock with price \$15 is worth \$5. If the option is not exercised, the exercise price must be above the price of the underlying and the value of the option is \$0. These observations can be summarized by saying that value, or "fair" price, at expiration is given by

$$V^c = \max\{S - E, 0\}, \tag{14.1}$$

where the "max" operator means that whichever is the larger (or the maximum) of 0 and $S - E$ is selected. Hence if $S - E = 5$ then $\max\{5, 0\} = 5$ and if $S - E = -2$ then $\max\{-2, 0\} = 0$. The formula for the value of a call option at expiry is graphed in Figure 14.1. The value is initially 0 until the point at which $S = E$. After this point, each additional dollar increase in stock price leads to a dollar increase in value.

Example 154 *On June 26 2003 GlaxoSmithKline stock was trading at \$41. The exercise prices for the option contracts directly above and below this price were \$40 and \$42.50. The table displays the value at expiry for these contracts for a selection of prices of GlaxoSmithKline stock at the expiration date.*

S	37.50	40	41	42.50	45	47.50
$\max\{S - 40, 0\}$	0	0	1	2.50	5	7.50
$\max\{S - 42.5, 0\}$	0	0	0	0	2.50	5

Setting aside the issue of timing of payments (formally, assuming that no discounting is applied) the profit, Π^c, from holding a call option is given by

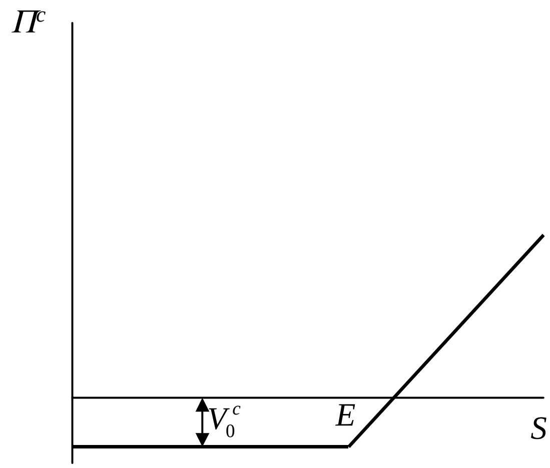


Figure 14.2: Profit from a Call Option

its value less the premium paid. If the premium is denoted V_0^c, profit can be written as

$$\begin{aligned}\Pi^c &= V^c - V_0^c = \max\{S - E, 0\} - V_0^c \\ &= \max\{S - E - V_0^c, -V_0^c\}. \qquad (14.2)\end{aligned}$$

The relationship between profit and the price of the underlying asset is graphed in Figure 14.2. The figure shows how the profit from purchasing a call option is potentially unlimited.

If a put option is exercised the holder receives a sum equal to the difference between the exercise price and the price of the underlying stock at expiry. An option with an exercise price of \$10 on an underlying stock with price \$5 is worth \$5. If the option is not exercised, the exercise price must be below the price of the underlying and the value of the option is \$0. These observations can be summarized by saying that value, or "fair" price, at expiration is given by

$$V^p = \max\{E - S, 0\}, \qquad (14.3)$$

so that the value is whichever is larger of 0 and $E - S$. The formula for the value of a put option is graphed in Figure 14.3. When the underlying stock price is 0, the option has value equal to the exercise price. The value then declines as the underlying price rises, until it reaches 0 at $S = E$. It remains zero beyond this point.

Example 155 *Shares in Fox Entertainment Group Inc. traded at \$29.72 on 7 July 2003. The expiry value of put options with exercise prices of \$27.50 and \$30.00 are given in the table for a range of prices for Fox Entertainment Group Inc. stock.*

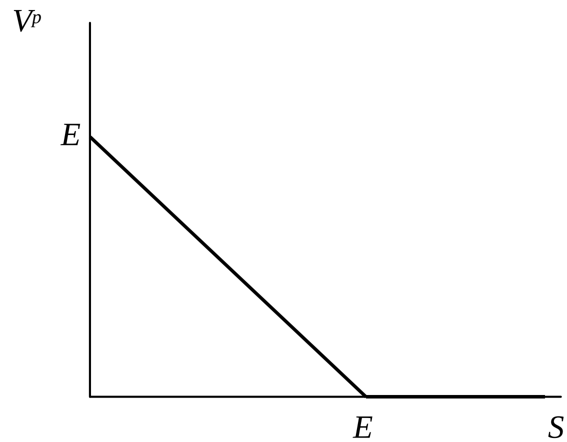


Figure 14.3: Value of a Put Option

S	20	22.50	25.00	27.50	30	32.50
$\max\{27.50 - S, 0\}$	7.50	5	2.50	0	0	0
$\max\{30 - S, 0\}$	10	7.50	5	2.50	0	0

The profit from purchasing a put option (again assuming no discounting so the timing of payments can be ignored) is given by the difference between the premium paid, V_0^p, and its value at expiry. Hence

$$\begin{aligned} \Pi^p &= V^p - V_0^p = \max\{E - S, 0\} - V_0^p \\ &= \max\{-V_0^p, E - S - V_0^p\}. \end{aligned} \tag{14.4}$$

This profit is graphed in Figure 14.4 as a function of the price of the underlying stock at expiry. The figure shows how the maximum profit from a put is limited to $E - V_0^p$.

These results can be extended to portfolios involving options. Consider a portfolio consisting of a_s units of the underlying stock, a_c call options and a_p put options. A short position in any one of the three securities is represented by a negative holding. At the expiry date, the value of the portfolio is given by

$$P = a_s S + a_c \max\{S - E, 0\} + a_p \max\{E - S, 0\}. \tag{14.5}$$

The profit from the portfolio is its final value less the purchase cost.

Example 156 *Consider buying two call options and selling one put option, with all options having an exercise price of \$50. If calls trade for \$5 and puts for \$10, the profit from this portfolio is*

$$\Pi = 2\max\{S - 50, 0\} - \max\{50 - S, 0\} - 2 \times 5 + 10.$$

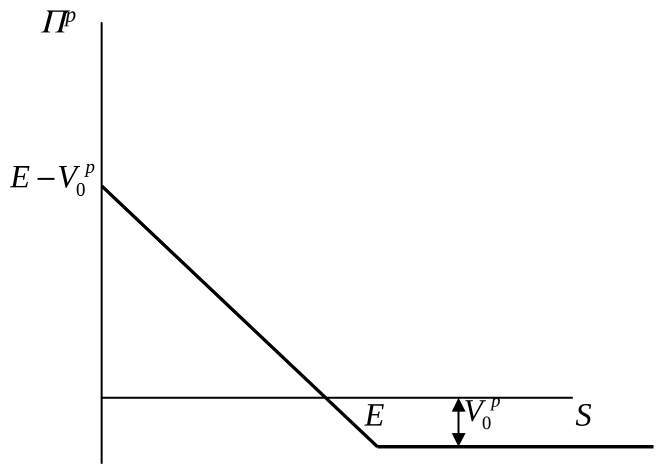


Figure 14.4: Profit from a Put Option

For $S < 50$*, the put will be exercised but not the call, leading to a profit of*

$$\Pi = -50 + S.$$

For $S > 50$*, the two calls are exercised, but not the put. The profit becomes*

$$\Pi = 2S - 100.$$

Profit is 0 *when* $S = 50$*.*

Portfolios of puts, calls and the underlying asset can be used to engineer different structures of payoffs. Several of these have been given colorful names representing their appearance. The first example is the *straddle* which involves buying a put and a call on the same stock. If these have the same exercise price, the profit obtained is

$$\Pi^P = \max\{E - S, 0\} + \max\{S - E, 0\} - V_0^p - V_0^c \tag{14.6}$$

The level of profit as a function of the underlying stock price is graphed in Figure 14.5. This strategy is profitable provided the stock price deviates sufficiently above or below the exercise price.

The *strangle* is a generalization of the straddle in which a put and call are purchased that have different exercise prices. Denoting the exercise price of the put by E^p and the that of the call by E^c, the profit of the strategy when $E^p < E^c$ is shown in Figure 14.6.

Finally, a *butterfly spread* is a portfolio constructed by purchasing a call with exercise price E_1^c and a call with exercise price E_3^c. In addition, two calls with exercise price E_2^c halfway between E_1^c and E_3^c are sold. The profit level is

$$\Pi^P = \max\{S - E_1^c, 0\} - 2\max\{S - E_2^c, 0\} + \max\{S - E_3^c, 0\} - V_1^c + 2V_2^c - V_3^c. \tag{14.7}$$

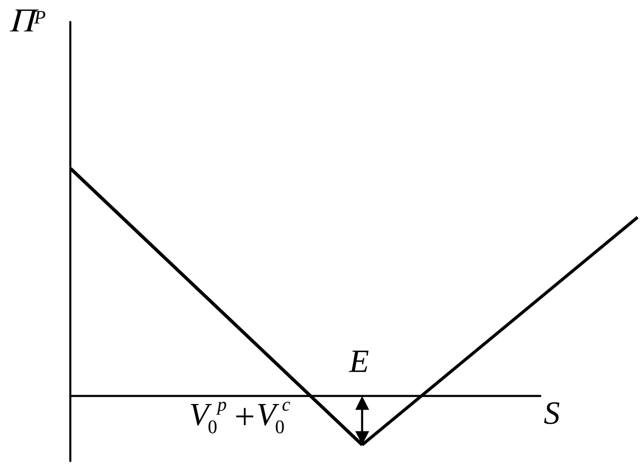


Figure 14.5: Profit from a Straddle

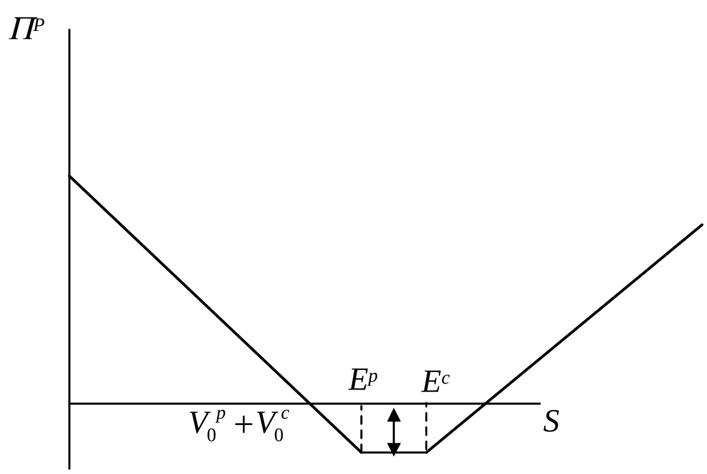


Figure 14.6: Profit from a Strangle

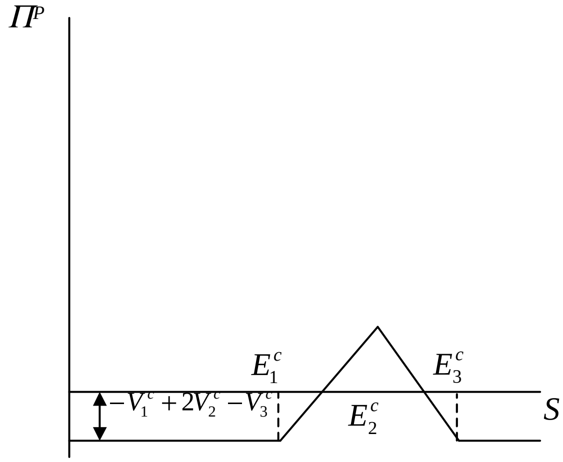


Figure 14.7: Profit from a Butterfly Spread

When the underlying stock price is below E_1^c none of the options is exercise and a profit is made equal to $-V_1^c + 2V_2^c - V_3^c$. For S between E_1^c and E_2^c profit rises. Once E_2^c has been reached, further increases in S reduce profit until E_3^c. Beyond this point, profit is again equal to $-V_1^c + 2V_2^c - V_3^c$. This is graphed in Figure 14.7.

14.4 Put-Call Parity

There is a relationship between the value of a call option and the value of a put option. In fact, if one value is known, the other can be derived directly. This relationship is determined by analyzing a particular portfolio of call, put and the underlying asset.

Consider a portfolio that consists of holding one unit of the underlying asset, one put option on that asset, and the sale of one call option, with the put and call having the same exercise price. If V^p is the value of the put option and V^c the value of the call, the value of the portfolio, P, is

$$P = S + V^p - V^c. \tag{14.8}$$

At the expiration date, the final values for the two options can be used to write the portfolio value as

$$P = S + \max\{E - S, 0\} - \max\{S - E, 0\}. \tag{14.9}$$

If $S < E$ at the expiration date, then the put option is exercised but not the call. The value of the portfolio is

$$P = S + E - S = E. \tag{14.10}$$

Conversely, if $S > E$ the call options is exercised but not the put. This gives the value of the portfolio as

$$P = S - S + E = E. \tag{14.11}$$

Hence, whatever the price of the underlying asset at the expiration date, the value of the portfolio is

$$P = E, \tag{14.12}$$

so the portfolio has the same value whatever happens to the stock price.

Since the value of the portfolio is constant for all S, the portfolio is a safe asset and must pay the rate of return earned on the risk-free asset. If this return is r, with continuous compounding the initial value of the portfolio if there are t units of time until the date of expiry is equal to the discounted value of the exercise price, so

$$S + V^p - V^c = Ee^{-rt}. \tag{14.13}$$

Therefore, at any time up to the expiration date, if either V^p or V^c is known, the other can be derived directly. This relationship is known as *put-call parity.*

Example 157 *A call option on a stock has 9 months to expiry. It currently trades for \$5. If the exercise price is \$45 and the current price of the underlying stock is \$40, the value of a put option on the stock with exercise price \$45 and 9 months to expiry when the risk-free rate is 5% is*

$$V^p = V^c - S + Ee^{-r[T-t]} = 5 - 40 + 45e^{-0.05\times 0.75} = 8.44.$$

14.5 Valuing European Options

The problem faced in pricing an option before the expiration date is that we do not know what the price of the underlying asset will be on the date the option expires. In order to value an option before expiry it is necessary to add some additional information. The additional information that we use takes the form of a model of asset price movements. The model that is chosen will affect the calculated price of the option so it is necessary work towards a model that is consistent with the observed behavior of asset prices.

The initial model that is considered makes very specific assumptions upon how the price of the underlying asset may move. These assumptions may seem to be too artificial to make the model useful. Ultimately though, they form the foundation for a very general and widely applied formula for option pricing.

The method of valuation is based on arbitrage arguments. The analysis of Arbitrage Pricing Theory emphasized the force of applying the idea that two assets with the same return must trade at the same price to eliminate arbitrage opportunities. To apply this to the valuation problem the process is to construct a portfolio, with the option to be valued as one of the assets in the portfolio, in such a way that the portfolio has the same return as an asset with known price. In essence, the returns on the portfolio are matched to the returns on another asset. The portfolio must then trade at the same price as the asset whose returns it matches. Knowing the prices of all the components of the portfolio except for the option then implies we can infer the value of the option. This simple methodology provides exceptionally powerful for valuing options and will be used repeatedly in what follows.

The analysis given in this section is for European options on an underlying stock that does not pay any dividends. Dividends can be incorporated using the same methodology but space limitations prevent this extension being undertaken here. The valuation of American options requires a development of the analysis for European options and is analyzed in Section 14.7.

14.5.1 The Basic Binomial Model

To begin the study of option pricing we first consider the very simplest model for which the valuation problem has any substance. Although simple, solving this teaches us all we need to know to progress to a very general solution.

Assume that when the option is purchased there is a single period to the expiration date. No restriction needs to be placed on the length of this period, as long as the rates of returns are defined appropriately for that period. When the contract is purchased, the current price of the underlying stock is known. What we do not know is the price of the underlying at the expiration date. If we did, we could calculate the profit from the option, discount it back to the date at which the contract is purchased and determine a precise value. It is this missing piece of information about the future price of the underlying stock that we must model. The modelling consists of providing a statistical distribution for the possible prices at the expiration date.

The fundamental assumption of the basic binomial model is that the price of the stock may take one of two values at the expiration date. Letting the initial price of the underlying stock be S, then the binomial assumption is that the price at the expiration date will either be:

- Equal to uS, an outcome which occurs with probability p;

or

- Equal to dS, an outcome which occurs with probability $1 - p$.

The labelling of these two events is chosen so that $u > d \geq 0$, meaning that the final price uS is greater than the price dS. Consequently, the occurrence of the price uS can be called the "good" or "up" state and price dS the "bad" or "down" state. It can be seen how this model captures the idea that the price of the underlying stock at expiration is unknown when the option is purchased.

The final component of the model is to assume that a risk-free asset with return r is also available. Defining the gross return, R, on the risk-free asset by

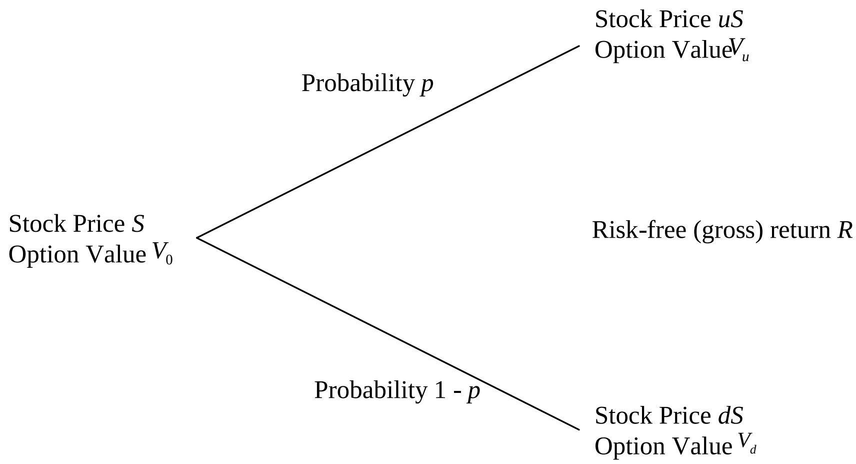


Figure 14.8: Binomial Tree for Option Pricing

$R \equiv 1 + r$, it must be case that the return on the risk-free asset satisfies

$$u > R > d. \tag{14.14}$$

This must hold since if $R > u$ the risk-free asset would always provide a higher return that the underlying stock. Since the stock is risky, this implies that no-one would hold the stock. Similarly, if $d > R$ no-one would hold the risk-free asset. In either of these cases, arbitrage possibilities would arise.

Section 14.3 has already shown how to value options at the expiration date. For example, if the stock price raises to uS, the value of a call option is $\max\{uS - E, 0\}$ and that of a put option is $\max\{E - uS, 0\}$. For the present, it is enough to observe that we can calculate the value of the option at the expiration date given the price of the underlying. The value of the option at expiration is denoted V_u when the underlying stock price is uS and V_d when it is dS.

The information that has been described can be summarized in *binomial tree diagram*. Consider Figure 14.8. At the left of the diagram is the date the option is purchased – denoted time 0. At this time the underlying stock price is S and the option has value V_0. It is this value V_0 that is to be calculated. The upper branch of the tree represents the outcome when the underlying price is uS at expiration and the lower branch when it is dS. We also note the risk-free return on the tree.

To use this model for valuation, note that there are three assets available: (1) the underlying stock; (2) the option; and (3) the risk-free asset. Constructing a portfolio of any two of these assets which has the same return as the third allows the application of the arbitrage argument.

Consequently, consider a portfolio that consists of one option and $-\Delta$ units of the underlying stock. The number of units of the underlying stock is chosen so that this portfolio has the same value when the underlying has price uS at the expiration date as it does when it has price dS. This then allows us to apply the arbitrage argument since the portfolio has a fixed value and so must pay the same return as the risk-free asset. The portfolio constructed in this way is often referred to as the "delta hedge" for the option.

The cost of this portfolio at the date the option contract is purchased is

$$P_0 = V_0 - \Delta S, \tag{14.15}$$

where V_0 is the unknown which is to be determined. At the expiration date the value of the portfolio is either

$$P_u = V_u - \Delta uS, \tag{14.16}$$

or

$$P_d = V_d - \Delta dS. \tag{14.17}$$

The value of Δ is chosen to ensure a constant value for the portfolio at the expiration date. Hence Δ must satisfy

$$V_u - \Delta uS = V_d - \Delta dS, \tag{14.18}$$

giving

$$\Delta = \frac{V_u - V_d}{S\left[u - d\right]}. \tag{14.19}$$

Substituting this value of Δ back into (14.16) and (14.17),

$$P_u = P_d = \frac{uV_d - dV_u}{u - d}, \tag{14.20}$$

so it does give a constant value as required.

The arbitrage argument can now be applied. The portfolio of one option and $-\Delta$ units of the underlying stock provides a constant return. Therefore it is equivalent to holding a risk-free asset. Given this, it must pay the same return as the risk-free or else one could be arbitraged against the other. Hence the gross return on the portfolio must be R which implies

$$P_u = P_d = RP_0. \tag{14.21}$$

Now substituting for P_0 and P_u gives

$$\frac{uV_d - dV_u}{u - d} = R\left[V_0 - \Delta S\right]. \tag{14.22}$$

Using the solution for Δ and then solving for V_0

$$V_0 = \frac{1}{R}\left[\frac{R - d}{u - d}V_u + \frac{u - R}{u - d}V_d\right]. \tag{14.23}$$

This result gives the fair value for the option that eliminates arbitrage opportunities. In an efficient market, this would be the premium charged for the option.

The valuation formula is defined for general values of V_u and V_d. What distinguishes calls and puts are the specific forms that these values take. These can be called the *boundary values*. These boundary values were calculated in Section 14.3.

Example 158 *For a call option with exercise price be E, the value of the option at the expiration date is either $V_u = \max\{uS - E, 0\}$ or $V_d = \max\{dS - E, 0\}$. The initial value of the call option is then*

$$V_0 = \frac{1}{R}\left[\frac{R-d}{u-d}\max\{uS - E, 0\} + \frac{u-R}{u-d}\max\{dS - E, 0\}\right].$$

Example 159 *Consider a call option with exercise price \$50 written on a stock with initial price \$40. The price of the underlying stock may rise to \$60 or to \$45 and the gross return on the risk-free asset is 115%. These values imply $u = 1.5, d = 1.125$ and $R = 1.15$. The value of the option at the expiration date is either $V_u = \max\{uS - E, 0\} = \max\{60 - 50, 0\} = 10$ or $V_d = \max\{dS - E, 0\}$ $= \max\{45 - 50, 0\} = 0$. The initial value of the call option is then*

$$V_0 = \frac{1}{1.15}\left[\frac{0.025}{0.375}10 + \frac{0.35}{0.375}0\right] = \$0.58.$$

Example 160 *For a put option with exercise price E, the value of the option at the expiration date is either $V_u = \max\{E - uS, 0\}$ or $V_d = \max\{E - dS, 0\}$. The initial value of the put option is then*

$$V_0 = \frac{1}{R}\left[\frac{R-d}{u-d}\max\{E - uS, 0\} + \frac{u-R}{u-d}\max\{E - dS, 0\}\right].$$

Example 161 *Consider a put option with exercise price \$50 written on a stock with initial price \$40. The price of the underlying stock may rise to \$60 or to \$45 and the gross return on the risk-free asset is 115%. These values imply $u = 1.5, d = 1.125$ and $R = 1.15$. The value of the option at the expiration date is either $V_u = \max\{E - uS, 0\} = \max\{50 - 60, 0\} = 0$ or $V_d = \max\{E - dS, 0\}$ $= \max\{50 - 45, 0\} = 5$. The initial value of the put option is then*

$$V_0 = \frac{1}{1.15}\left[\frac{0.025}{0.375}0 + \frac{0.35}{0.375}5\right] = \$4.06.$$

The valuation formula we have constructed can be taken in two ways. On one level, it is possible to just accept it, and the more general variants that follow, as a means of calculating the value of an option. Without developing any further understanding they can be used to provide fair values for options that can then be applied in investment analysis. At a second level, the structure of the formula can be investigated to understand why it comes out the way it does and what the individual terms mean. Doing so provides a general method of valuation that can be applied to all valuation problems.

To proceed with the second approach, observe that the weights applied to V_u and V_d in (14.23) satisfy

$$\frac{R-d}{u-d} > 0, \frac{u-R}{u-d} > 0, \tag{14.24}$$

and

$$\frac{u-R}{u-d} + \frac{R-d}{u-d} = 1. \tag{14.25}$$

Since both weights are positive and their sum is equal to 1, they have the basic features of probabilities. To emphasize this, define $q \equiv \left[\frac{R-d}{u-d}\right]$. Then the valuation formula can be written as

$$V_0 = \frac{1}{R}\left[qV_u + [1-q]\,V_d\right]. \tag{14.26}$$

In this expression, the value V_0 is found by calculating the expected value at expiration and discounting back to the initial date using the risk-free rate of return. This shows that the value can be written in short form as

$$V_0 = \frac{1}{R}E_q\left(V\right), \tag{14.27}$$

where the subscript on the expectation operator indicates that the expectation is taken with respect to the probabilities $\{q, 1-q\}$.

The idea that we value something by finding its expected value in the future and then discount this back to the present is immediately appealing. This is exactly how we would operate if we were risk-neutral. However, the assumption in models of finance is that the market is on average risk-averse so that we cannot find values this simply. How this is captured in the valuation formula (14.27) is that the expectation is formed with the probabilities $\{q, 1-q\}$ which we have constructed *not* the true probabilities $\{p, 1-p\}$. In fact, the deviation of $\{q, 1-q\}$ from $\{p, 1-p\}$ captures the average risk aversion in the market. For this reason, the probabilities $\{q, 1-q\}$ are known as a *risk-neutral probabilities* – they modify the probabilities so that we can value *as if* we were risk-neutral.

This leaves open two questions. Firstly, where do the true probabilities $\{p, 1-p\}$ feature in the analysis? So far it does not appear that they do. The answer to this question is that the true probabilities are responsible for determining the price of the underlying stock. Observe that the price of the underlying stock when the option is purchased must be determined by its expected future payoffs. Hence, S is determined from uS and dS by a combination of the probabilities of the outcomes occurring, $\{p, 1-p\}$, discounting, R, and the attitude to risk of the market. The true probability may be hidden, but it is there.

Secondly, are these risk-neutral probabilities unique to the option to be valued? The answer to this question is a resounding *no*. When risk-neutral probabilities can be found they can be used to value all assets. In this analysis there are only three assets but all can be valued by using the risk neutral probabilities. Consider the underlying stock. For this asset, $V_u = uS$ and $V_d = dS$. Using these in the valuation formula

$$V_0 = \frac{1}{R}\left[qV_u + [1-q]V_d\right] = \frac{1}{R}\left[\frac{R-d}{u-d}uS + \frac{u-R}{u-d}dS\right] = S. \tag{14.28}$$

Hence the risk neutral probabilities also value the underlying stock correctly. For the risk-free asset

$$V_0 = \frac{1}{R}\left[qV_u + [1-q]V_d\right] = \frac{1}{R}\left[\frac{R-d}{u-d}R + \frac{u-R}{u-d}R\right] = 1. \tag{14.29}$$

This process of calculating he expected value of returns using the risk-neutral probabilities and then discounting back to the present using the risk-free rate of return is therefore a general valuation method that can be applied to all assets.

Example 162 *Consider a call option with exercise price \$50 written on a stock with initial price \$50. The price of the underlying stock may rise to \$60 or fall to \$45 and the gross return on the risk-free asset is 110%. The risk-neutral probabilities are given by*

$$q = \frac{R-d}{u-d} = \frac{1.1-0.9}{1.2-0.9} = \frac{2}{3}, \tag{14.30}$$

and

$$1-q = \frac{u-R}{u-d} = \frac{1.2-1.1}{1.2-0.9} = \frac{1}{3}. \tag{14.31}$$

The initial value of the call option is then

$$V_0 = \frac{1}{R}E_q(V) = \frac{1}{1.1}\left[\frac{2}{3}10 + \frac{1}{3}0\right] = \$6.06.$$

In addition, the price of the underlying stock must satisfy

$$V_0 = \frac{1}{1.1}\left[\frac{2}{3}60 + \frac{1}{3}45\right] = \$50.$$

14.5.2 The Two-Period Binomial

The single-period binomial model introduced a methodology for valuing options but does not represent a very credible scenario. Where it is lacking is that the underlying stock will have more than two possible final prices. Having introduced the method of risk-neutral valuation, the task of relaxing this restriction and moving to a more convincing environment is not at all difficult.

A wider range of final prices can be obtained by breaking the time period between purchase of the option and the expiration date into smaller sub-intervals and allowing the stock price to undergo a change over each sub-interval. As long as the rate of return for the risk-free asset and the proportional changes in the

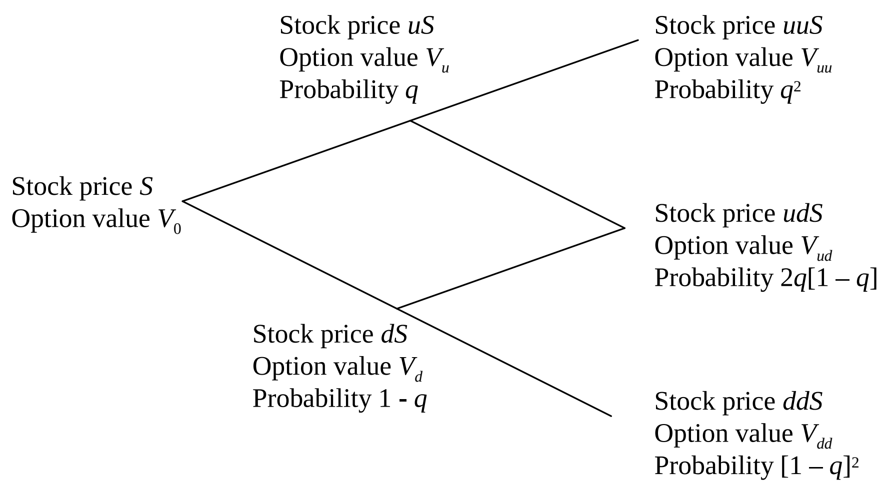


Figure 14.9: Binomial with Two Sub-Intervals

stock price are defined relative to the length of each sub-interval, the use of risk-neutral valuation can be directly extended to this setting.

Consider Figure 14.9 which shows the period between purchase and expiration broken into two sub-intervals. Starting with an underlying stock price of S, at the end of the first sub-interval the price will either be uS or dS. In terms of the risk-neutral probabilities, these will occur with probabilities q and $1-q$ respectively. Starting from the price uS, it is possible to reach a final price at the end of the second interval of either uuS or udS. Since the probability of another u is q and of a d is $1-q$, these final prices must have probabilities q^2 and $q\,[1-q]$ respectively. Similarly, starting from dS, another d occurs with probability $1-q$ and a u with probability q. Hence the final price ddS is reached with probability $[1-q]^2$ and duS with probability $[1-q]\,q$. But $udS = duS$, so the central price at the expiration date can be reached in two different ways with total probability of arrival given by $2q\,[1-q]$. The risk-free (gross) return, R, is defined as the return over each sub-interval. Hence the return over the period is R^2. This completes the construction of the figure.

The value of the option V_0 can be obtained in two different ways. The first way is to use a two-step procedure which employs risk-neutral valuation to find V_u and V_d using the expiration values, and then uses these to find V_0. Although not strictly necessary for a European option, it is worth working through these two steps since this method is necessary when American options are valued. The second way to value the option is to apply risk-neutral valuation directly to the expiration values using the compound probabilities. Both give the same answer.

To apply the two-step procedure, assume we are at the end of the first sub-interval. The price of the underlying stock is either uS or dS. If it is uS, then applying (14.26) the value of the option must be

$$V_u = \frac{1}{R}\left[qV_{uu} + [1-q]\,V_{ud}\right]. \tag{14.32}$$

Similarly, if the price of the underlying stock at the end of the first sub-interval is dS, the value of the option is

$$V_d = \frac{1}{R}\left[qV_{ud} + [1-q]\,V_{dd}\right]. \tag{14.33}$$

Now move to the very beginning of the tree. At the end of the first sub-interval the option is worth either V_u or V_d. Applying risk-neutral valuation, the value of the option at the purchase date must be

$$V_d = \frac{1}{R}\left[qV_u + [1-q]\,V_d\right]. \tag{14.34}$$

Substituting into this expression using (14.32) and (14.33) gives

$$V_0 = \frac{1}{R^2}\left[q^2V_{uu} + 2q\,[1-q]\,V_{ud} + [1-q]^2\,V_{dd}\right]. \tag{14.35}$$

This is the fair value of the option at the purchase date. It should also be clear that this is the result that would have been obtained by applying risk-neutral valuation directly to the values at the expiration date using the risk-neutral probabilities given in the binomial tree.

Example 163 *For a call option with exercise price E,*

$$\begin{aligned} V_{uu} &= \max\left\{u^2S - E, 0\right\}, \\ V_{ud} &= \max\left\{udS - E, 0\right\}, \\ V_{dd} &= \max\left\{d^2S - E, 0\right\}. \end{aligned}$$

The value of the call is

$$\begin{aligned} V_0 &= \frac{1}{R^2}\left[q^2V_{uu} + 2q\,[1-q]\,V_{ud} + [1-q]^2\,V_{dd}\right] \\ &= \frac{1}{R^2}\left[\begin{array}{c} q^2\max\left\{u^2S - E, 0\right\} + 2q\,[1-q]\max\left\{udS - E, 0\right\} \\ +\,[1-q]^2\max\left\{d^2S - E, 0\right\} \end{array}\right]. \end{aligned}$$

Example 164 *Consider a put option with a year to expiry on a stock with initial price of \$50. Over a six month interval the stock can rise by 15% or by 5% and the risk-free rate of return is 107.5%. If the put option has an exercise price of \$65 the value of the contract is*

$$V_0 = \frac{1}{[1.075]^2}\left[\frac{1}{4}^2 0 + \frac{1}{4}\frac{3}{4}4.625 + \frac{3}{4}^2 9.875\right] = \$5.557.$$

14.5.3 The General Binomial

The process of either working back through the tree or applying risk-neutral valuation directly to the expiration values can be applied to a binomial tree with any number of sub-intervals. A general variant of the binomial formula is now obtained that applies whatever number of sub-intervals the period is divided into.

To derive this, note that in (14.35) the occurrence of a q in the expression matches the occurrence of a u (and a $1-q$ matches a d). Furthermore, the coefficients on the values at expiration are 1, 1 for the one-interval case and 1, 2, 1 for the two-interval case. These are the terms in the standard binomial expansion. Using these observations, the valuation formula for a period divided into n sub-intervals can be immediately derived as

$$V_0 = \frac{1}{R^n}\left[\sum_{j=0}^{n}\left[\frac{n!}{j!\,[n-j]!}\right] q^j\,[1-q]^{n-j}\,V_{u^j d^{n-j}}\right]. \tag{14.36}$$

It is easy to check that for $n = 1$ and $n = 2$ this gives the results already derived directly.

Example 165 *When $n = 4$ the valuation formula is*

$$V_0 = \frac{1}{R^4}\left[\begin{array}{c} q^4 V_{u^4} + 4q^3\,[1-q]\,V_{u^3 d} + 6q^2\,[1-q]^2\,V_{u^2 d^2} \\ +4q\,[1-q]^3\,V_{ud^3} + [1-q]^4\,V_{d^4} \end{array}\right].$$

Although the result as given is time consuming to use when computing results manually, it is easy to write a program that will compute it automatically. Even when n is large, it will only take seconds to obtain an answer. The valuation formula is therefore perfectly usable and the next sub-section shows that it can reflect the actual data on stock price movements. However, it can be improved further by specifying the details of the final valuations.

Consider a call option. In this case

$$V_{u^j d^{n-j}} = \max\left\{u^j d^{n-j} S - E, 0\right\}. \tag{14.37}$$

Now define a as the smallest non-negative integer for which

$$u^a d^{n-a} S - E > 0. \tag{14.38}$$

Hence it requires a minimum of a "up" moves to ensure that the option will be in the money at expiry. Consequently, if $j < a$ then $\max\left\{u^j d^{n-j} S - E, 0\right\} = 0$ and if $j > a$ then $\max\left\{u^j d^{n-j} S - E, 0\right\} = u^j d^{n-j} S - E$. With this definition of a it is only necessary to include the summation in (14.36) the terms for $j \geq a$, since all those for lower values of a are zero.

Example 166 *Let $u = 1.05, d = 1.025, S = 20, E = 24$ and $n = 5$. The values of $u^j d^{n-j} S$ are given in the table.*

u^0d^5S	u^1d^4S	u^2d^3S	u^3d^2S	u^4d^1S	u^5d^0S
22.63	23.18	23.75	24.32	24.92	25.53

Example 167 *It can be seen that $u^jd^{n-j}S$ exceeds the exercise price E of 24 only when $j \geq 3$. Hence $a = 3$.*

Using the definition of a to remove from the summation those outcomes for which the option is worthless at expiry, the value of the call becomes

$$V_0 = \frac{1}{R^n}\left[\sum_{j=a}^{n}\left[\frac{n!}{j!\,[n-j]!}\right]q^j\,[1-q]^{n-j}\left[u^jd^{n-j}S - E\right]\right]. \tag{14.39}$$

Separating this expression into terms in S and terms in E,

$$V_0 = S\left[\sum_{j=a}^{n}\left[\frac{n!}{j!\,[n-j]!}\right]q^j\,[1-q]^{n-j}\left[\frac{u^jd^{n-j}}{R^n}\right]\right] \tag{14.40}$$

$$-ER^{-n}\left[\sum_{j=a}^{n}\left[\frac{n!}{j!\,[n-j]!}\right]q^j\,[1-q]^{n-j}\right]. \tag{14.41}$$

Now define

$$q' = \frac{u}{R}q, 1 - q' = \frac{d}{R}[1-q], \tag{14.42}$$

and let

$$\Phi(a;n,q) \equiv \left[\sum_{j=a}^{n}\left[\frac{n!}{j!\,[n-j]!}\right]q^j\,[1-q]^{n-j}\right], \tag{14.43}$$

and

$$\Phi(a;n,q') \equiv \left[\sum_{j=a}^{n}\left[\frac{n!}{j!\,[n-j]!}\right]\left[\frac{uq}{R}\right]^j\left[\frac{d\,[1-q]}{R}\right]^{n-j}\right]. \tag{14.44}$$

$\Phi(a;n,q)$ (and equivalently $\Phi(a;n,q')$) is the *complementary binomial distribution function* which gives the probability that the sum of n random variables, each with value 0 with probability q and value 1 with probability $1-q$ will be greater than or equal to a. Because they are probabilities, both $\Phi(a;n,q)$ and $\Phi(a;n,q')$ must lie in the range 0 to 1.

Using this notation, the valuation formula can be written in the compact form

$$V_0 = \Phi(a;n,q')\,S - ER^{-n}\Phi(a;n,q). \tag{14.45}$$

The value of the option is therefore a combination of the underlying stock price and the discounted value of the exercise price with each weighted by a probability. This is an exceptionally simple formula.

14.5.4 Matching to Data

The next question to be addressed is how to make the formula in (14.45) into a result that can be applied in a practical context. To evaluate the formula we need to supply values for S, E, R, n and q. The underlying stock price S and the risk-free return R can be obtained directly from market data. The exercise price E is written into the option contract. The number of intervals, n, is chosen to trade-off accuracy against ease of computation. All that is unknown is q, the probability in the binomial tree.

To motivate the approach taken to providing a value for q, recollect that the basic idea of the binomial tree is that the price of the underlying stock is random. Given a value of R, the value of q is determined by u and d. The values of u and d must be chosen to result in behavior of the underlying stock price that mirrors that observed in the market place. This leads to the idea of fixing u and d to provide a return and variance of the underlying stock price in the binomial model that equals the observed variance of the stock in market data.

Let the observed expected return on the stock be $\overline{r}$ and its variance be σ^2. Each of these is defined over the standard period of time. If the time length of each interval in the binomial tree is Δt, the expected return and variance on the stock over an interval are $\overline{r}\Delta t$ and $\sigma^2\Delta t$. If at the start of an interval the stock price is S, the expected price at the end of the interval using the observed return is $Se^{\overline{r}\Delta t}$. Matching this to the expected price in the binomial model gives

$$puS + [1-p]\,dS = Se^{\overline{r}\Delta t}, \tag{14.46}$$

where it should be noted that these are the probabilities of the movements in the statistical model, not the risk-neutral probabilities. Solving this shows that to match the data

$$p = \frac{e^{\overline{r}\Delta t} - d}{u - d}. \tag{14.47}$$

Over an interval in the binomial tree, the return on the underlying stock is $u-1$ with probability p and $d-1$ with probability $1-p$. The expected return is therefore $pu + [1-p]\,d - 1$. The variance in the binomial model, σ_b^s, can then be calculated as $\sigma_b^s = pu^2 + [1-p]\,d^2 - [pu + [1-p]\,d]^2$. Equating this variance is to the observed market variance gives

$$pu^2 + [1-p]\,d^2 - [pu + [1-p]\,d]^2 = \sigma^2\Delta t. \tag{14.48}$$

Substituting for p from(14.47), ignoring terms involving powers of Δt^2 and higher, a solution of the resulting equation is

$$u = e^{\sigma\sqrt{\Delta t}}, \tag{14.49}$$

$$d = e^{-\sigma\sqrt{\Delta t}}. \tag{14.50}$$

These values can then be used to parameterize the binomial model to match observed market data.

Example 168 *The data in Example 36 generate an annual variance of 523.4% for General Motors stock. If the year is broken into 365 intervals of 1 day each, then* $\Delta t = \frac{1}{365} = 0.00274$ *and* $\sigma = 22.88$*. Hence*

$$u = e^{\sigma\sqrt{\Delta t}} = e^{22.88\sqrt{0.00274}} = 3.3, \tag{14.51}$$

and

$$d = e^{-\sigma\sqrt{\Delta t}} = e^{-22.88\sqrt{0.00274}} = 0.3. \tag{14.52}$$

These imply

$$q = \frac{e^{\overline{r}\Delta t} - d}{u - d} = \frac{e^{6.5\times 0.00274} - 0.3}{3.3 - 0.3} = 0.239. \tag{14.53}$$

14.6 Black-Scholes Formula

In moving from the single-interval binomial to the general binomial the process used was to reduce the interval between successive price changes. Continuing to shorten the interval eventually leads to a situation where one price change follows another without any time seeming to have passed. In the limit, we can then think of price changes occurring continuously, rather than at discrete intervals as in the binomial. Such continuity comes close to capturing the observation that for most significant stocks a very large number of trades take place so the actual process is almost continuous.

Taking the limit of the binomial model as the interval between trades shrinks to zero leads to the *Black-Scholes equation.* The Black-Scholes equation is one of the most fundamental results in investment analysis. Its value comes from the fact that it provides an easily applied practical solution to the problem of pricing options that can be evaluated using observable market data. The construction of the equation revolutionized the way option markets functioned since it provided an exact and easily computable fair value for an option.

The move from discrete intervals in the binomial model to continuous time for Black-Scholes leads to two changes to the valuation formula (14.45). The first is very simple: the discrete compounding captured in the term R^{-n} becomes the continuous analog e^{-rT} where T is the time until the option expires and r is the risk-free interest rate for a compatible time period. For example, if the option has 9 months to expiry and r is the annual risk-free rate then T is defined as written as a fraction of a year, in this case $T = 0.75$.

The second change relates to the probabilities. In the general binomial formula S and E are weighted by values from complementary binomial distributions. In the limit as the length of the time intervals shrink to zero, these distributions converge to the normal distribution and the weights become values from the cumulative function for the normal distribution. Being the cumulative of the normal distribution, both weights are again between 0 and 1.

Collecting these points together, the Black-Scholes equation for the value of a call option is given by

$$V^c = N(d_1)\,S - Ee^{-rT}N(d_2)\,, \tag{14.54}$$

where $N(d_1)$ and $N(d_2)$ are values from the cumulative normal distribution and

$$d_1 = \frac{\ln(S/E) + \left[r + 0.5\sigma^2\right]T}{\sigma\sqrt{T}}, \tag{14.55}$$

$$d_2 = \frac{\ln(S/E) + \left[r - 0.5\sigma^2\right]T}{\sigma\sqrt{T}}. \tag{14.56}$$

Recalling the discussion of applying the general binomial formula, S, E, r, T can be directly observed and σ calculated from observed market data. Given these values, the formula is applied by computing d_1 and d_2 then determining $N(d_1)$ and $N(d_2)$ from statistical tables for the cumulative normal – a table is contained in the appendix. The formula is then evaluated.

Example 169 *A call option with an exercise price of \$40 has three months to expiry. The risk-free interest rate is 5% per year and the stock price is currently \$36. If the standard deviation of the asset price is 0.5, then $T = 0.25$, $E = 40$, $S = 36$, $\sigma = 0.5$ and $r = 0.05$. The formulas for the call option give*

$$d_1 = \frac{\ln(36/40) + \left[0.05 + 0.5(0.5)^2\right]0.25}{0.5\sqrt{0.25}} = -0.25,$$

and

$$d_2 = \frac{\ln(36/40) + \left[0.05 - 0.5(0.5)^2\right]0.25}{0.5\sqrt{0.25}} = -0.5.$$

From the tables for the cumulative normal distribution

$$N(d_1) = 0.4013, N(d_2) = 0.3083.$$

Substituting into the Black-Scholes formula

$$V^c = [0.4013 \times 36] - \left[\frac{40}{e^{0.05\times0.25}} \times 0.3085\right] = \$2.26.$$

The Black-Scholes formula for the value for a put of option is

$$V^p = N(-d_2)\,Ee^{-rT} - N(-d_1)\,S, \tag{14.57}$$

where the definitions of d_1 and d_2 are as for a call option.

Example 170 *If $T = 0.25$, $E = 40$, $S = 36$, $\sigma = 0.5$ and $r = 0.05$ then $d_1 = -0.25$ and $d_2 = -0.5$. From the cumulative normal tables*

$$N(-d_1) = N(0.25) = 0.5987,$$

and

$$N(-d_2) = N(0.5) = 0.695.$$

This gives the value of the put as

$$V^p = \left[0.695 \times \frac{40}{e^{0.05\times0.25}}\right] - [0.5987 \times 36] = \$5.90.$$

14.7 American Options

The analysis of European options is much simplified by the fact that they can only be exercised at the expiration date. The fact that American options can be exercised at any time up until the date of expiry adds an additional dimension to the analysis. It now becomes necessary to determine the best time to exercise.

The best way to analyze this is to return to the two-interval binomial tree displayed in Figure 14.9. The two-interval model provides a time after the first price change at which the issue of early exercise can be addressed. With American options it is also necessary to treat calls and puts separately.

14.7.1 Call Options

Assume that a call option is being analyzed and that the first price change has lead to a price of uS for the underlying stock. The holder of the option then has three choices open to them:

- Exercise the option and obtain $\max\{uS - E, 0\}$;
- Hold the option and receive either V_{uu}^c or V_{ud}^c depending on the next price change;
- Sell the option for its value V_u^c.

Whether the option should be exercised depends on which of these three alternatives leads to the highest return. First consider holding the option. The payoff of this strategy can be evaluated by employing risk-neutral valuation. Hence the value of receiving either V_{uu}^c or V_{ud}^c is

$$V_u^c = \frac{1}{R}\left[qV_{uu}^c + [1-q]\,V_{ud}^c\right], \tag{14.58}$$

but this is precisely the fair market value of the option. The value of holding the option is therefore the same as that of selling (though there is risk involved in the former). Now compare exercising to selling. If the option is sold, V_u^c is realized. If it is exercised, $uS - E$ is realized - there is no point exercising the option if $uS - E < 0$. By definition

$$V_{uu}^c = \max\left\{u^2S - E, 0\right\} \geq u^2S - E, \tag{14.59}$$

and

$$V_{ud}^c = \max\{udS - E, 0\} \geq udS - E. \tag{14.60}$$

Using risk-neutral valuation and the inequalities in (14.59) and (14.60)

$$\begin{aligned} V_u^c &= \frac{1}{R}\left[qV_{uu}^c + [1-q]\,V_{ud}^c\right] \\ &\geq \frac{1}{R}\left[qu^2S + [1-q]\,udS - E\right]. \end{aligned} \tag{14.61}$$

But

$$\begin{aligned}\frac{1}{R}\left[qu^2S+[1-q]\,udS-E\right] &= \frac{u\,[qu+[1-q]\,d]\,S}{R}-\frac{E}{R} \\ &> uS-E, \end{aligned} \tag{14.62}$$

where the last inequality follows from the fact that $qu+[1-q]\,d=R$ and $R\geq 1$. Combining these statements

$$V_u^c > uS-E, \tag{14.63}$$

which shows that the option should never be exercised early. It is always better to hold or to sell than to exercise.

Similarly, if after the first interval the price is dS, the choice of strategies is:

- Exercise and obtain $\max\{dS-E,0\}$;
- Hold the option and receive either V_{ud}^c or V_{dd}^c depending on the next price change;
- Sell the option for its value V_d^c.

Applying risk-neutral valuation shows that the second and third provide the payoff V_d^c. Noting that

$$V_{dd}^c = \max\left\{d^2S-E,0\right\} \geq d^2S-E, \tag{14.64}$$

then

$$\begin{aligned} V_d^c &= \frac{1}{R}\left[qV_{du}^c+[1-q]\,V_{dd}^c\right] \\ &\geq \frac{1}{R}\left[qudS+[1-q]\,d^2S-E\right], \\ &= \frac{d\,[qu+[1-q]\,d]\,S}{R}-\frac{E}{R} \\ &> dS-E. \end{aligned} \tag{14.65}$$

Hence the conclusion obtained is that

$$V_d^c > dS-E, \tag{14.66}$$

so that it is better to hold or sell than to exercise.

These calculations illustrate the maxim that an option is "Better alive than dead", revealing that an American call option will never be exercised early. It is always better to hold or to sell than exercise. Even though the options have the feature of early exercise, if they are priced correctly this should never be done.

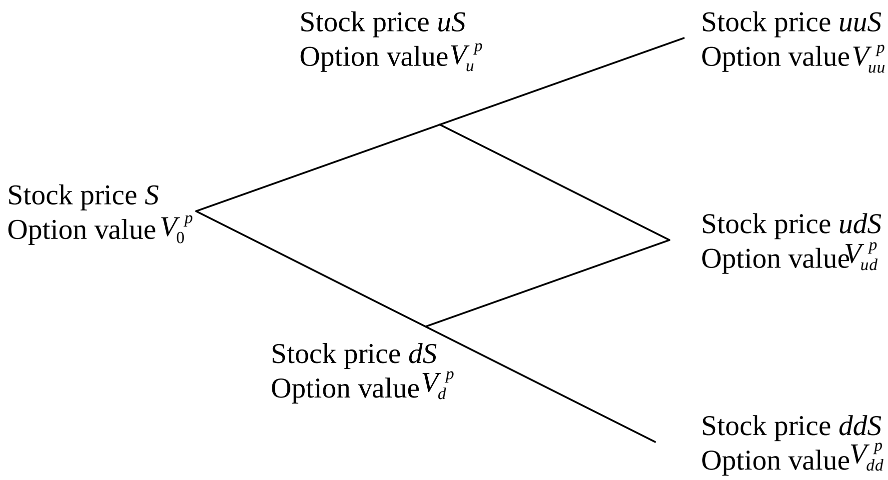


Figure 14.10: American Put Option

14.7.2 Put Option

The same conclusion cannot be obtained for a put option. In this case it may be better to exercise.

The two-interval binomial tree for an American put option is illustrated in Figure 14.10. Consider being at the end of the first interval and observing a stock price of dS. The value of the put option at this point is V_d^p and early exercise would obtain the amount $E - dS$. The option should therefore be exercised early if $E - dS > V_d^p$. Where an American put differs from an American call is that this can hold in some circumstances and early exercise becomes worthwhile.

This can be seen by using the expiration values and the risk-neutral probabilities to obtain

$$V_d^p = \frac{1}{R}\left[qV_{du}^p + [1-q]\, V_{dd}^p\right]. \tag{14.67}$$

Numerous possibilities now arise depending upon whether V_{du}^p and V_{dd}^p are positive or zero. That early exercise can be optimal is most easily demonstrated if both are taken to be positive. In this case, $V_{du}^p = E - udS$ and $V_{dd}^p = E - ddS$. Then early exercise will be optimal if

$$E - dS > \frac{1}{R}\left[q\left[E - udS\right] + [1-q]\left[E - ddS\right]\right]. \tag{14.68}$$

Substituting for q and $1 - q$ then solving shows that the inequality in (14.68) holds if

$$R > 1. \tag{14.69}$$

Therefore, the put option will be exercised early if the return on the risk-free asset is positive.

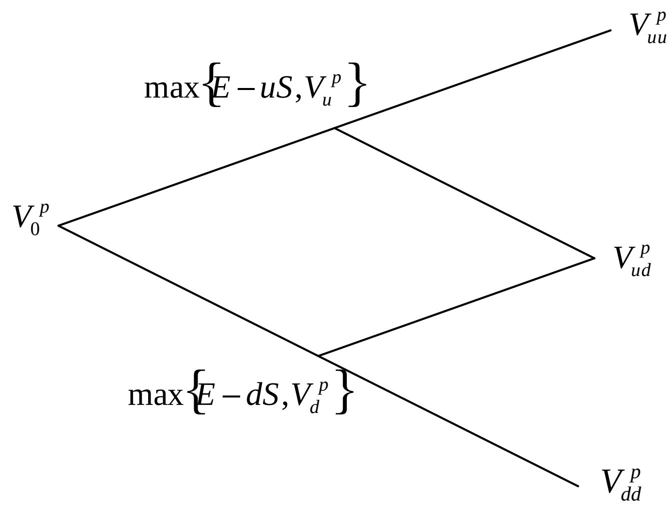


Figure 14.11: Value of American Put

A similar analysis can be undertaken to investigate the numerous other possibilities. But the important conclusion is that it is sometimes optimal to exercise American puts early. So their value must be higher than for a European put. The actual method of valuation of an American put option is to construct the binomial tree and to assign the value of the option at each node as the maximum of the early exercise value and the fair value of the option. This is shown in Figure 14.11 which indicates the value at each node incorporating the option for early exercise.

Example 171 *Consider a two-interval binomial tree with* $R = 1.05, u = 1.1, d = 1$ *and an initial stock price of \$10.*

A European put option contract with exercise price \$12 is worth

$$\begin{aligned} V_0^p &= \frac{1}{R^2}\left[q^2 V_{uu}^p + 2q\left[1-q\right]V_{ud}^p + \left[1-q\right]^2 V_{dd}^p\right] \\ &= \frac{1}{1.05^2}\left[0.25\times 0 + 0.5\times 1 + 0.25\times 2\right] \\ &= \$0.907. \end{aligned}$$

An American put on the same stock has value

$$V_0^p = \frac{1}{R}\left[q\max\left\{E-uS, V_u^p\right\} + \left[1-q\right]\max\left\{E-dS, V_d^p\right\}\right].$$

Working back from the end of the binomial tree,

$$\begin{aligned} V_u^p &= \frac{1}{R}\left[qV_{uu}^p + [1-q]\,V_{ud}^p\right] \\ &= \frac{1}{1.05}\left[0.5\times 0 + 0.5\times 1\right] \\ &= \$0.476, \end{aligned}$$

and

$$\begin{aligned} V_d^p &= \frac{1}{R}\left[qV_{ud}^p + [1-q]\,V_{dd}^p\right] \\ &= \frac{1}{1.05}\left[0.5\times 1 + 0.5\times 2\right] \\ &= \$1.429. \end{aligned}$$

Therefore $\max\{E-uS, V_u^p\} = \max\{12-11, 0.476\} = 1$ *(so the option is exercised early) and* $\max\{E-dS, V_d^p\} = \max\{12-10, 1.429\} = 2$ *(so the option is exercised early). The initial value of the option is then*

$$\begin{aligned} V_0^p &= \frac{1}{1.05}\left[0.5\times 1 + 0.5\times 2\right] \\ &= \$1.429. \end{aligned}$$

As claimed, if it is optimal to exercise early, the American option has a higher value than the European option.

14.8 Summary

The chapter has described call and put options, distinguishing between European and American contracts. Information on where these options can be traded and where price information can be found has also been given.

The process of valuing these options began with a determination of the value of the options at the expiration date. From these results the profit from portfolios of options was determined. In particular, this process was used to derive put-call parity.

It was then noted that to provide a value before the expiration date it was necessary to model the statistical distribution of future prices of the underlying stock. European options were valued using the single-period binomial model. The model was then gradually generalized, eventually resulting in the Black-Scholes formula.

American options were then considered. It was shown that an American call would never be exercised early but a put may be. American calls therefore have the same value as European calls. American puts will be at least as valuable as European puts and may be strictly more valuable.

14.9 Exercises

Exercise 97 *Consider two call options on the same underlying stock. Option 1 has an exercise price of \$60 and sells for \$5 while option 2 has an exercise price of \$55 and sells for \$6. Assuming they have the same expiration date, calculate the profit from the strategy of issuing two \$60 calls and purchasing one \$55. Sketch the level of profit versus the share price at the expiration date.*

Exercise 98 *If a call option on a stock trading at \$40 has an exercise price of \$45 and a premium of \$2, determine the premium on a put option with the same exercise price if the annual risk-free rate of return is 5% and there is 6 months to expiration.*

Exercise 99 *Using the binomial pricing model calculate the value of a call option on a stock that currently sells for \$100 but may rise to \$115 or fall to \$80 when there is 1 year to expiry, the risk free rate of return is 5% and the exercise price is \$105. Repeat this exercise breaking the year in (i) two six month intervals and (ii) three four month intervals but retaining \$115 and \$80 as the maximum and minimum prices that can be reached.*

Exercise 100 *Prove that (14.49) and (14.50) are a solution to the equation relating observed market variance to the variance in the binomial model.*

Exercise 101 *Taking the prices from Yahoo, find the sample variance for Ford stock (ten years of data) and hence compute u and d for a daily sub-interval.*

Exercise 102 *Determine the value of a call option with 9 months to go before expiration when the stock currently sells for \$95, has an instantaneous standard deviation of 0.8, the exercise price is \$100 and the continuously compounded risk-free rate of return in 6%.*

Exercise 103 *Consider a stock that currently trades for \$75. A put and call on this stock both have an exercise price of \$70 and expire in 150 days. If the risk-free rate is 9 percent and the standard deviation for the stock is 0.35, compute the price of the options using Black-Scholes.*

Exercise 104 *Show that the values given for put and call options satisfy put-call parity.*

Exercise 105 *Consider a two-interval binomial tree with $S = 20, E = 22, u = 1.1, d = 1.025$ and $R = 1.05$. By applying the two-step procedure to work back through the tree, show that an American call option on the stock will never be exercised early.*

Exercise 106 *For the data in Exercise 105, determine at which points in the tree the put will be exercised early.*

Chapter 15

Forwards and Futures

> *Please excuse the facts the notes are not perfect but all the material you need is here.*

15.1 Introduction

Forwards and futures are both contracts which involve the delivery of a specific asset at an agreed date in the future at a fixed price. They differ from options contracts in the fact that there is no choice involved as to whether the contract is exercised. With both forwards and futures the agreed price must be paid and delivery undertaken. Despite this, the underlying approach to valuation remains the same.

Forward contracts, which are no more than commitments to a future trade, have been in use for a very long time. One piece of evidence to this effect is that the agreement to purchase dates whilst the dates were still unripe on the tree (a forward contract) was prohibited in the early Islamic period. Commodity futures also have a fairly long history. They were first introduced onto an exchange by the Chicago Board of Trade in the 1860s to assist with the reduction in trading risk for the agriculatural industry. Financial futures (which differ in significant ways from commodity futures) are a much more recent innovation.

This chapter will introduce the main features of forward and future contracts and describe where they can be traded. The motives for trading and potential trading strategies will be analyzed. Finally, the valuation of the contracts will be considered.

15.2 Forwards and Futures

Forwards and futures are two variants of the same basic transaction but there are some important operational differences between them. These differences are reflected in the valuations of the contracts. The forward contract is the simpler form and this is described first.

As an example of a forward contract consider a farmer growing wheat and a baker who requires wheat as an ingredient. Assume that wheat is harvested in September. A forward contract would be written if the farmer and the baker committed in May to the baker purchasing 2 tons of wheat at $1000 per ton when the wheat is harvested in September. The essential elements here are the commitment to trade at a future date for a fixed price and quantity. No money is exchanged when the forward contract is agreed. Money only changes hands when the commodity is delivered. The financial question that arises is the determination of the price (in this example $1000) written into the contract.

A futures contract has the almost all the features of forward contract. In a futures contract there is also a commitment to trade and agreed quantity at a fixed price at a future date. Where differences arise between forwards and futures is in the timing of institutioanl arrangement and the timing of payment.

- A forward is an *over-the-counter* agreement between two individuals. In contrast, a future is a trade organized by an *exchange*.

- A forward is settled on the *delivery date*. That is, there is a single payment made when the contract is delivered. The profit or loss on a future is settled on a *daily basis*.

To understand the process of daily settlement, assume a futures contract is agreed for delivery of a commodity in three months. Label the day the contract is agreed as day 1 and the day of delivery as day 90. Let the delivery price written into the contract on day 1 be $30. Now assume that on day 2 new contracts for delivery on day 90 have a delivery price of $28 written into them. Those who are holding contracts with an agreement to pay $30 are in a worse position than than those holding $28 contracts. The daily settlement process requires them to pay $2 (the value by which their position has deteriorated) to those who have sold the contract. The delivery price of $28 on day 2 is then taken as the starting point for day 3. If the delivery price in new contracts rises to $29 on day 3 then the holder of the futures contract from day 2 receives $1. This process is repeated every day until day 90. Effectively, daily settlement involves the futures contract being re-written each day with a new contract price.

From this brief description, it can be seen that a futures contract involves a continuous flow of payments over the life of the contract. In contrast, a forward contract has a single payment at the end of the contract. This difference in the timing of payments implies that the contracts need not have the same financial valuation.

With a futures contract the exchange acts as an intermediary between the two parties on different sides of the contract. The process of daily settlement is designed to avoid the development of excessive negative positions and the possibility of default. To further reduce the chance of default exchanges insist upon the maintenance of margin. Margin must be maintained by both parties to a sufficient level to cover daily price changes.

The next section of the chapter will focus upon the trading details of futures contracts because these are the contracts that can be most readily traded. The focus will then shift to forward contracts when valuation is considered. The reason for focussing on forward contracts is that the single payment involved makes valuation a very much simpler process. Finally a contrast will be drawn between the valuation of a forward contract and the valuation of a futures contract.

15.3 Futures

There are two basic types of futures contracts. These are *commodity futures* and *financial futures*.

15.3.1 Commodity Futures

Commodity futures are trades in actual commodities. Many significant agricultural products are covered by futures contracts including wheat, pork and orange juice plus other commodities such as timber. Futures contracts originated in an organized way with the Chicago Board of Trade and have since been offered by numerous other exchanges.

Example 172 *The Chicago Board of Trade was established in 1848. It has more than 3,600 members who trade 50 different futures and options products through open auction and/or electronically. Volume at the exchange in 2003 was 454 million contracts. Initially, only agricultural commodities such as corn, wheat, oats and soybeans were traded. Futures contracts have developed to include non-storable agricultural commodities and non-agricultural products such as gold and silver. The first financial futures contract was launched in October 1975 based on Government National Mortgage Association mortgage-backed certificates. Since then further futures, including U.S. Treasury bonds and notes, stock indexes have been introduced. Options on futures were introduced in 1982.*

(http://www.cbot.com/cbot/pub/page/0,3181,1215,00.html)

A contract with the Board of Trade, which is similar in structure to contracts on other exchanges, specifies:

- The *quality* of the product. The quality has to be very carefully defined so that the parties to the contract know exactly what will be traded. This is important when there are many different varieties and qualities of the same product.

- The *quantity* of a trade. The quantity that is traded is specified in the contract. This is usually large in order to make delivery an economically viable exercise. However, it does mean that these contracts are "lumpy" so that the assumption of divisibility is not easily applied.

- The *place* to which delivery is made. The importance of this is through the high transport costs that can be involved in shipping the commodities around.
- The *date* of delivery (or interval in which delivery is to be made). This is essential for the contract to function.
- The *price*. This is the basic feature of the contract upon which profit and loss is determined. The price is what will be paid at the delivery time.

These specifications have to be very precise and complete in order to ensure that there can be no dispute about whether the correct product is ultimately delivered.

Example 173 ***Soybeans Futures.*** *1. Contract Size 5,000 bu. 2. Deliverable Grades No. 2 Yellow at par, No. 1 yellow at 6 cents per bushel over contract price and No. 3 yellow at 6 cents per bushel under contract price *No. 3 Yellow Soybeans are only deliverable when all factors equal U.S. No. 2 or better except foreign material. See Chapter 10s - Soybean Futures in the Rules & Regulations section. 3. Tick Size 1/4 cent/bu ($12.50/contract) 4. Price Quote Cents and quarter-cents/bu 5. Contract Months Sep, Nov, Jan, Mar, May, Jul, Aug 6. Last Trading Day The business day prior to the 15th calendar day of the contract month. 7. Last Delivery Day Second business day following the last trading day of the delivery month. 8. Trading Hours Open Auction: 9:30 a.m. - 1:15 p.m. Central Time, Mon-Fri. Electronic: 7:31 p.m. - 6:00 a.m. Central Time, Sun.-Fri. Trading in expiring contracts closes at noon on the last trading day. 9. Ticker Symbols Open Auction: S, Electronic: ZS 10. Daily Price Limit: 50 cents/bu ($2,500/contract) above or below the previous day's settlement price. No limit in the spot month (limits are lifted two business days before the spot month begins).*
(http://www.cbot.com/cbot/pub/cont_ detail/0,3206,959+14397,00.html)

Although the contracts specify delivery of a commodity, most contracts are closed before the delivery date. Less then 1% are delivered or settled in cash.

15.3.2 Financial Futures

Financial futures are contracts drawn up on the basis of some future price or index such as the interest rate or a stock index. Generally, no "good" is delivered at the completion of the contract and only a financial exchange takes place. Generally is used because there are exceptions involving bond contracts.

Financial futures become possible when it is observed that the actual commodity need not be delivered – at the end of the contract only the "profit" over the current spot price is paid. For example, assume the futures contract price is $3 and the spot price is $2. Then the buyer of futures contract pays $1 to the seller and no transfer of asset needs to take place.

A financial future can also be formed by converting an index into a monetary equivalent. For instance, a stock index future can be constructed by valuing each

10 points at \$1. Thus an index of 6100 would trade at a price of \$610. If the index fell to 6000, the futures price would become \$60. Using such a mechanism, it becomes possible to construct such contracts on any future price.

Example of exchanges in the US where financial futures are traded are the Chicago Board of Trade, Mid-America Commodity Exchange and New York Board of Trade.

Example 174 *NYSE Composite Index® Futures*

Contract REVISED NYSE Composite Index® Futures Small Contract Size \$5 × NYSE Composite Index (e.g., \$5 × 5000.00 = \$25,000) Symbol Value of Minimum Move MU \$2.50

Contract REVISED NYSE Composite Index® Futures Reg. Contract Size \$50 × NYSE Composite Index® (e.g., \$50 × 5000.00 = \$250,000) Symbol Value of Minimum Move YU \$25.00

Price Quotation: Index Points where 0.01 equals \$0.50

Daily Price Limits: Please contact the Exchange for information on daily price limits for these contracts.

Position Limits: NYSE Regular (on a 10:1 basis) are converted into NYSE Small positions for limit calculation purposes. Any One Month Limit 20,000 All Months Combined Limit 20,000

Cash Settlement: Final settlement is based upon a special calculation of the third Friday's opening prices of all the stocks listed in the NYSE Composite Index® .

(http://www.nybot.com/specs/yxrevised.htm)

In the UK, futures contracts are traded on LIFFE – the London International Financial Futures Exchange – which was opened in 1982.

Example 175 *LIFFE offers a range of futures and options, and provides an arena for them to be traded. The Exchange brings together different parties – such as financial institutions, corporate treasury departments and commercial investors, as well as private individuals – some of whom want to offset risk, hedgers, and others who are prepared to take on risk in the search for profit.*

Following mergers with the London Traded Options Market (LTOM) in 1992 and with the London Commodity Exchange (LCE) in 1996, LIFFE added equity options and a range of soft and agricultural commodity products to its existing financial portfolio. Trading on LIFFE was originally conducted by what's known as "open outcry". Traders would physically meet in the Exchange building to transact their business. Each product was traded in a designated area called a pit, where traders would stand and shout the price at which they were willing to buy or sell.

In 1998, LIFFE embarked on a programme to transfer all its contracts from this traditional method of trading, to an electronic platform. This transition is now complete. The distribution of LIFFE CONNECTTM stands at around 450 sites, more than any other trading system in the world, and covers all major time zones. This distribution continues to grow.

(http://liffe.npsl.co.uk/liffe/site/learning.acds?instanceid=101765&context=100190)

There are three major types of futures traded on LIFFE.

- *Contracts on short term interest rates* These are based on the three-month money market rate and are priced as 100 - interest rate. Consequently, when the interest rate goes up it implies the price of the futures contract goes down.
- *Bond futures* Bond future represent long-term interest rate futures. They are settled by delivery of bonds, with adjustment factors to take account of the range of different bonds that may be delivered. This is a financial future which is settled by actual delivery of the commodity.
- *Equity index futures* Equity index futures are cash settled and are priced per index point.

Foreign Exchange Quotes

Futures exchange rates are quoted as the number of USD per unit of the foreign currency

Forward exchange rates are quoted in the same way as spot exchange rates.

This means that GBP, EUR, AUD, and NZD are USD per unit of foreign currency.

Other currencies (e.g., CAD and JPY) are quoted as units of the foreign currency per USD.

15.4 Motives for trading

Two motives can be identified for trading forwards and futures. These are *hedging* and *speculation*. These motives are now discussed in turn.

15.4.1 Hedging

Hedging is the use of the contracts to reduce risk. Risk can arise from either taking demanding or supplying a commodity at some time in the future. The current price is known but the price at the time of demand or supply will not be known. A strategy of hedging can be used to guard against unfavorable movements in the product price.

Two examples of the way in which hedging can be employed are now given.

Example 176 *Consider a bakery which needs wheat in three months. It can:*

i. wait to buy on the spot market;

or

ii. buy a future now.

If the baker followed (ii) they would be a long hedger – this is the investor who has committed to accept delivery.

Example 177 *Consider a company in the UK who will be paid in three months time in Euros. It can:*

i. sell a future on the Euros now;

or

ii. wait to receive the Euros and sell them on the spot market.

If the firm followed (i) they would be a short hedger – the investor who commits to supply the commodity.

The advantage of a futures contract is that it fixes the price and guards against price changes. For someone who has to buy in the future it can be used to insure against price increases while for someone who has to sell in the future it can insure against price falls.

A company that is due to sell an asset at a particular time in the future can hedge by taking a short futures position. They then hold a *short hedge*. A company that is due to buy an asset at a particular time in the future can hedge by taking a long futures position – a *long hedge*.

Arguments in Favor of Hedging

Companies should focus on the main business they are in and take steps to minimize risks arising from interest rates, exchange rates, and other market variables

Arguments against Hedging

Shareholders are usually well diversified and can make their own hedging decisions

It may increase risk to hedge when competitors do not

Explaining a situation where there is a loss on the hedge and a gain on the underlying can be difficult

Basis Risk

Bais is the difference between spot and futures

The basis at time t is

$$\text{Basis} = S_t - F_t$$

Basis risk arises because of the uncertainty about the basis when the hedge is closed out

The basis will change after the hedge is constructed

Suppose that

F_1 : Initial Futures Price

F_2 : Final Futures Price

S_2 : Final Asset Price

1. Short Hedge

The future sale of an asset can be hedged by entering into a short futures contract

Assume that the position is closed at time 2 and the asset sold at the spot price S_2

The profit on the future is $F_1 - F_2$

$$\begin{aligned} \text{Price Realized} &= S_2 + (F_1 - F_2) \\ &= F_1 + \text{Basis} \end{aligned}$$

2. Long Hedge

The future purchase of an asset can be hedged by entering into a long futures contract

Enter hedge at F_1, close at F_2, purchase at S_2

The loss on the hedge is $F_1 - F_2$

$$\begin{aligned} \text{Cost of Asset} &= S_2 + (F_1 - F_2) \\ &= F_1 + \text{Basis} \end{aligned}$$

Choice of Contract

Choose a delivery month that is as close as possible to, but later than, the end of the life of the hedge

When there is no futures contract on the asset being hedged, choose the contract whose futures price is most highly correlated with the asset price.

There are then two components to basis

$$\text{Price} = F_1 + (S_2^* - F_2) + (S_2 - S_2^*)$$

Optimal Hedge Ratio

Hedging through the use of futures contracts reduces risk by fixing a delivery or purchase price. This insures against adverse price movements but also means that profit is lost from favorable price movements. The optimal degree of hedging determines the best trade-off between these. In effect, it is usually best to cover some exposure by hedging but leave some uncovered in order to profit from favorable price movements. The *hedge ratio* is the size of the position in futures relative to size of exposure

One way of analyzing the optimal degree of hedging is to consider the strategy that minimizes the variance in a position. The optimal hedge ratio can be determined by considering the variation in the spot price and the futures price.

Let ΔS be change in spot price S over length of hedge and ΔF be change in futures price F over length of hedge. The standard deviation of ΔS is denoted by σ_S and the standard deviation of ΔF by σ_F. Let ρ be coefficient of correlation between ΔS and $\Delta F$and let the hedge ratio be denoted by h.

Consider a position which is long in the asset but short in future. With h denoting the hedge ratio, the change in the value of the position over the life of the hedge is

$$\Delta P = \Delta S - h\Delta F. \tag{15.1}$$

Conversely, when long in the future but short in the asset the change in value of position is

$$\Delta P = h\Delta F - \Delta S. \tag{15.2}$$

For both of these positions, the variance of change in the value of hedged position is

$$\begin{aligned} var(\Delta P) &= E\left(\Delta P - E\left(\Delta P\right)\right)^2 \\ &= E\left(\Delta S - h\Delta F - E\left(\Delta S - h\Delta F\right)\right)^2. \end{aligned} \tag{15.3}$$

Computing the expectation gives

$$var(\Delta P) = \sigma_S^2 + h^2\sigma_F^2 - 2h\rho\sigma_S\sigma_F. \tag{15.4}$$

One definition of an optimal policy is to choose the hedge ratio to minimize this variance. The necessary condition for the hedge ratio is

$$\frac{dvar(\Delta P)}{dh} = 2h\sigma_F^2 - 2\rho\sigma_S\sigma_F = 0. \tag{15.5}$$

Solving this condition, the hedge ratio that minimizes the variance is

$$h = \rho\frac{\sigma_S}{\sigma_F}. \tag{15.6}$$

Given data on these standard deviations and the correlation, this optimal hedge ratio is simple to compute.

Example 178 *A company must buy 1m gallons of aircraft oil in 3 months. The standard deviation of the oil price is 0.032. The company hedges by buying futures contracts on heating oil. The standard deviation is 0.04 and the correlation coefficient is 0.8. The optimal hedge ratio is*

$$0.8 \times \frac{0.032}{0.040} = 0.64.$$

One heating oil futures contract is for 42000 gallons. The company should buy

$$0.64 \times \frac{1000000}{42000} = 15.2,$$

contracts, which is 15 when rounded.

The example illustrates that the hedge does not have to be in the same commodity but only in a similar commodity whose price is highly correlated with the one being hedged. In addition, it also show that optimal hedging does not necessarily imply that all of the exposure has to be covered. In the example the company has an exposure of 1m gallons but buys futures contracts of 630000 gallons.

15.4.2 Stock Index Futures

A stock index tracks a hypothetical portfolio.
Some examples of contracts on stock indexes are:

- Dow Jones Industrial Average (30 US blue-chip) – CBOT \$10 times index, \$5 times index
- S&P 500 (400 industrials, 40 utilities, 20 transport, 40 financial) – CME \$250 times index, \$50 times index
- Nikkei 225 Stock Average (largest stock) – CME \$5 times index

Contracts on stock indexes are settled in cash.
A stock index future can hedge an equity portfolio.
Let P = current value of portfolio
Let A = current value of stocks underlying one future contract
If the portfolio mirrors the index then the optimal hedge ratio $h = 1$ and

$$N^* = P/A$$

For example, if index = 1000, contract is \$250 times index, $P = 1,000,000$ and $A = 250,000$ then $N^* = 4$

The portfolio will rarely mirror the index
The relative variability then needs to be taken into account
Recall β from the CAPM model
β measure of riskiness of a portfolio
If $\beta > 1$ the portfolio is more risky than the market
The number of futures contracts that should be shorted is

$$N^* = \beta P/A$$

This result is illustrated in the next example.
Value of S&P 500 $= 1,000$
Value of portfolio $= \$5,000,000$
Beta of portfolio $= 1.5$
Use a four month contract to hedge for three months
Current futures price is $1,010$
One contract is for \$250 times the index, so

$$A = 250 \times 1,000 = 250,000$$

This gives the optimal number of contracts to short as

$$N^* = 1.5 \times 5,000,000/250,000 = 30$$

Now the effect of the hedge can be assessed.
Assume in three months index = 900

futures price = 902
Risk-free interest rate = 4% (per annum)
Dividend yield = 1% (per annum)
Gain from short futures position

$$30 \times (1,010 - 902)x250 = \$810,000$$

The index declines by 10% over the 3 months. The dividend yield is 1% per annum, so is 0.25% over 3 months. This gives a return of −9.75% for the index. The expected return on the portfolio can be computed using the Security Market Line from the Capital Asset Pricing Model. The hedge is for 3 months so the risk-free rate is 1%. The expected return on the portfolio for three months is

$$1.0 + [1.5x(-9.75 - 1.0)] = -15.125$$

This gives the expected value of portfolio as

$$\$5,000,000x(1 - 0.15125) = \$4,243,750$$

The expected value of hedged position is

$$\$4,243,750 + \$810,000 = \$5,053,750$$

Reasons for Hedging an Equity Portfolio
Desire to be out of the market for a short period of time.
Hedging may be cheaper than selling the portfolio and buying it back.
Desire to hedge systematic risk
Appropriate when you feel that you have picked stocks that will outperform the market.

Changing Beta
The previous formula is for completely hedging the portfolio
This gives the hedged portfolio a $\beta = 1$
Hedging a different amount allows a chosen beta to be achieved
To change β to β^* (with $\beta > \beta^*$) adopt a short position in

$$(\beta - \beta^*)P/A$$

contracts
To change β to β^* (with $\beta < \beta*$) adopt a long position in

$$(\beta^* - \beta)P/A$$

contracts

Rolling The Hedge Forward
A series of futures contracts can be used to increase the life of a hedge
Each time a switch is made from one futures contract to another a type of basis risk is incurred.

15.4.3 Speculation

The second reason for trading in futures is speculation. If the spot price is expected to change, a trader can engage in speculation through futures.

A speculator has no interest in taking delivery of the commodity or of supplying it, but is simply interested in obtaining profit through trade. Consequently, any trade they make must ultimately be matched by a reversing trade to ensure that they do not need to receive or deliver.

For an expected price rise a speculator will:

i. Buy futures now;

ii. Enter a reversing trade to sell later after the price has risen.

Conversely, for an expected price fall, the speculator will:

i. Sell futures now;

ii. Enter a reversing trade to buy later after the price has fallen.

Clearly, even though the quantity of commodity to be traded is limited to the amount produced, any number of speculative trades can be supported if there are speculators on both sides of the market.

15.5 Forward Prices

The valuation issue involved with forward contracts is to determine the delivery price, or forward price, that is written into the contract at its outset. At the time the two parties on either side of a contract agree the trade, no payment is made. Instead the forward price is set so that the contract is “fair” for both parties. To be fair the contract must have a value of zero at the time it is agreed. It is this fact that allows the delivery price to be determined.

As we will see, the forward price in the contract and the spot price of the underlying asset at the time the contract is agreed are related. This relationship is now developed as the basis for determining the forward price.

This section develops the valuation of forward contracts. Forwards are considered since the daily settlement involved in futures contracts makes their analysis more complex. A later section explores the extent of the differences between the values of the two contracts.

The focus of this section is upon investment assets. The important feature of these is that it is possible to go short in these assets or reduce a positive holding if it is advantageous to do so. This allows us the flexibility to apply an arbitrage argument to obtain the forward price. A number of cases are considered which differ in whether or not the asset pays an income.

15.5.1 Investment Asset with No Income

The process of valuation using arbitrage involves searching for profitable opportunities by combining the assets that are available. To determine the fair futures price it is assumed that the assets available consist of a risk-free asset, the asset underlying the forward contract and the forward contract. If the forward price is not correctly set, it becomes possible to produce arbitrage profits by combining these assets.

The construction of an arbitrage portfolio is illustrated by the following example.

Example 179 *Consider a stock with a current spot price of $40, which will pay no dividends over the next year, and a one-year risk free rate of 5%. Suppose that the forward price for delivery in one year is $45 and a contract is for 100 shares. Given these numbers, it is possible to earn an arbitrage profit.*

To achieve the profit, the following investment strategy is used:

1. *Borrow $4000 for 1 year at the interest rate of 5%;*
2. *Buy 100 shares of the stock for $4000;*
3. *Enter into a forward contract to sell 100 shares for $4500 in 1 year.*

On the delivery date of the forward contract at the end of 1 year, the loan requires $\$4000e^{0.05} = \$4205.1$ *to repay. The stock is sold for $4500. Hence a profit of $294.9 is earned. Note that this profit is entirely certain since all agreements are made at the outset of the forward contract. In particular, it does not depend on the price of the underlying stock at the delivery date. Since a risk-free profit can be earned, the forward price of 45 cannot be an equilibrium.*

Now consider the formulation of an investment strategy for a lower forward price.

Example 180 *Consider a stock with a current spot price of $40, which will pay no dividends over the next year, and a one-year risk free rate of 5%. Suppose that the forward price for delivery in one year is $40 and a contract is for 100 shares. Given these numbers, it is possible to earn an arbitrage profit.*

To achieve the profit, the following investment strategy is used:

1. *Sell short 100 shares of the stock for $4000;*
2. *Lend $4000 for 1 year at the interest rate of 5%;*
3. *Enter into a forward contract to buy 100 shares for $4000 in 1 year.*

On the delivery date of the forward contract at the end of 1 year, the loan is repaid and provides an income of $\$4000e^{0.05} = \$4205.1$*. The stock is purchased for $4000. Hence a profit of $205.1 is earned. This profit is entirely certain so the forward price of 40 cannot be an equilibrium.*

In the first example, the loan requires $4205.1 to repay, so no profit will be earned if the sale at the forward price earns precisely this same amount. Similarly, in the second example, no profit is earned if the purchase of the shares costs $4205.1. Putting these observations together, the only forward price that eliminates arbitrage profits has to be $42.05. This price satisfies the relation that

$$42.05 = 40e^{0.05}. \tag{15.7}$$

That is, the forward price is the current spot price compounded at the risk-free rate up to the delivery date.

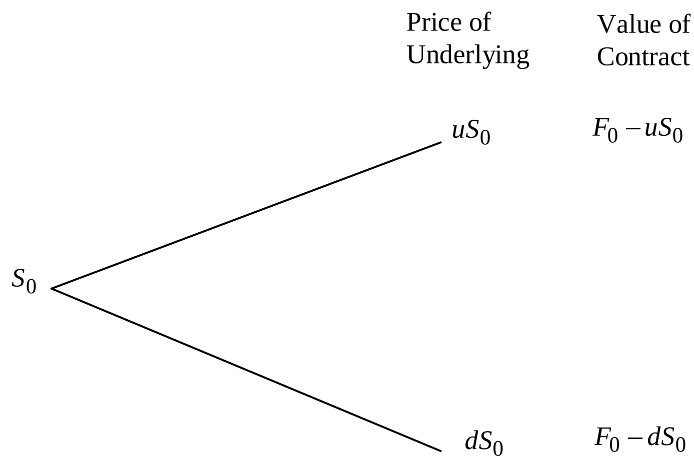


Figure 15.1: Binomial Tree for a Forward Contract

To express this for a general forward contract on an investment asset with no dividend, let the forward price at the outset of the contract be F_0, the spot price be S_0, the continuously compounded risk-free interest rate be r and the time to the delivery date be T. The forward price agreed at the outset of the contract must then be

$$F_0 = S_0 e^{rT}. \tag{15.8}$$

The construction of an arbitrage portfolio is only one method of obtaining the forward price. Recall that a similar process lead to the valuation of an option in the binomial model. Approaching forward contracts from this second direction emphasizes the generality of the method of valuation and shows that futures are not distinct from options.

Consequently, assume that spot price of the underlying asset at the outset of the contract is S_0. Adopting the binomial assumption, the price of the underlying stock can change to either uS_0 or dS_0 at the delivery date in the forward contract. For the investor who is short in the contract, the value of the forward contract at the delivery date is either $F_0 - uS_0$ when the asset price is uS_0 or $F_0 - dS_0$ when the price is dS_0. These prices and values produce the binomial tree in Figure 15.1.

Risk-neutral valuation can now be applied to the binomial tree. Let the risk-neutral probability associated with a move to uS_0 be q and that with a move to dS_0 be $1 - q$. With an option contract, a premium is paid for the contract and it is the fair value of this that is determined by risk-neutral valuation. In contrast, with a forward contract no payment is made or received at the start of the contract. Instead, the price in the contract F_0 is chosen to make the contract "fair", or to give it zero initial value. Letting V_0^f be the initial value of a futures contract, then F_0 must satisfy

$$V_0^f = \frac{1}{R}\left[q\left[F_0 - uS_0\right] + (1-q)\left[F_0 - dS_0\right]\right] = 0. \tag{15.9}$$

Solving this equation for F_0

$$F_0 = quS_0 + (1-q)\, dS_0. \tag{15.10}$$

Using the fact that $q = \frac{R-d}{u-d}$, this can be simplified to

$$F_0 = RS_0, \tag{15.11}$$

which is precisely the same price as in (15.8) when expressed in terms of discrete compounding.

Furthermore, for a binomial tree with n sub-periods, the initial forward price can be shown to satisfy

$$F_0 = R^n S_0, \tag{15.12}$$

so it converges to the result with continuous discounting as $n \to \infty$. Hence, risk-neutral valuation in the binomial tree can be used to value forward contracts in exactly the same way as for options.

15.5.2 Investment Asset with Known Income

Many financial assets provide an income to the holder. The holder of a forward on the asset does not receive this income, but the price of the underlying asset decreases to reflect the payment of the income. This observation allows the payment of income to be incorporated into the binomial tree.

If the asset pays an income with present value of I just prior to the delivery date in the forward contract, the value of the asset will be reduced to $uS_0 - IR$ on the upper branch of the tree and $dS_0 - IR$ on the lower branch. The modified binomial tree is in Figure 15.2.

The application of risk-neutral valuation gives

$$V_0^f = \frac{1}{R}\left[q\left[F_0 - uS_0 + IR\right] + (1-q)\left[F_0 - dS_0 + IR\right]\right] = 0. \tag{15.13}$$

Solving this using the definitions of the risk-neutral probabilities provides the forward price

$$F_0 = \left[S_0 - I\right] R. \tag{15.14}$$

As before, this can be extended naturally to the continuous case as

$$F_0 = \left[S_0 - I\right] e^{rT}. \tag{15.15}$$

Therefore, if the asset pays an income this reduces the forward price because the person who is long in the forward contract does not receive this income but is affected by the fall in the assets price immediately after the income is paid.

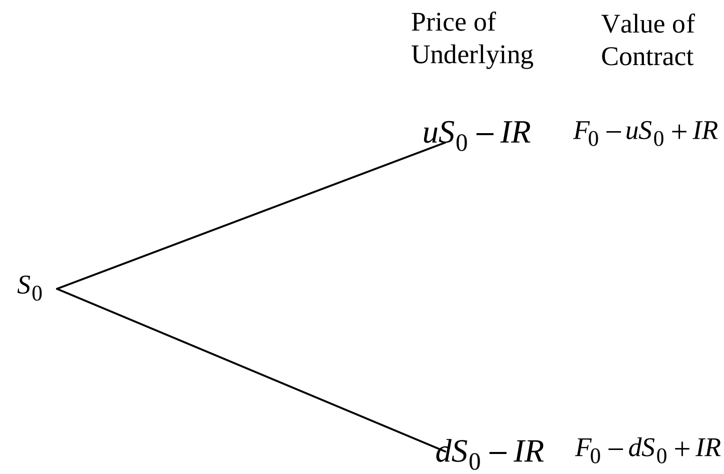


Figure 15.2: An Asset with Income

15.5.3 Continuous Dividend Yield

Rather than making a single payment of income, an asset may have a continuous flow of dividends. Let the rate of flow of dividends be q. Then the previous result can be modified to

$$F_0 = S_0 e^{(r-q)T}. \tag{15.16}$$

A continuous flow of dividends has the effect of continually reducing the asset price so reduces the forward price.

15.5.4 Storage costs

Storage costs are the opposite of income. They can be added into the expressions directly.

Let U be present value of storage costs then

$$F_0 = [S_0 + U]\, e^{rT}. \tag{15.17}$$

15.6 Value of Contract

It has already been noted that at the outset of the contract the forward price is chosen to ensure that the value of the contract is zero. As time progresses, the spot price of the underlying asset will change as will the forward price in new contracts. The contract can then either have a positive value if the price change moves in its favor and negative if it moves against.

To determine this value, let F_t be forward price at time t, and F_0 the forward price in a contract agreed at time 0. THIS CALCULATES the value using the argument that a sure sum of $F_t - F_0$ is delivered at time T. This is discounted back to t to give the value in the expression. SO it is the value at time t. Need to make this fact clear.

With time $T-t$ to the delivery date, the value, V_t^f, of the forward contract is then given by

$$V_t^f = [F_t - F_0]\, e^{-r[T-t]}. \tag{15.18}$$

As already noted, at the time the contract is written its value is zero. Now since $F_t = S_t e^{r[T-t]}$ it follows that the value of the contract at time t is

$$V_t^f = S_t - F_0 e^{-r[T-t]}. \tag{15.19}$$

With an income from the asset, this value becomes

$$V_t^f = S_t - I - F_0 e^{-r[T-t]}. \tag{15.20}$$

With a dividend at rate q the value of the forward at time t it is $F_t = S_t e^{(r-q)[T-t]}$. Substituting into (15.18) gives

$$V_t^f = \left[S_t e^{(r-q)[T-t]} - F_0\right] e^{-r[T-t]}. \tag{15.21}$$

Simplifying this expression

$$V_t^f = S_t e^{-q[T-t]} - F_0 e^{-r[T-t]}. \tag{15.22}$$

Finally, using the fact that $F_0 = S_0 e^{(r-q)T}$

$$V_t^f = S_t e^{-q[T-t]} - S_0 e^{-qT+rt}. \tag{15.23}$$

15.7 Commodities

Considering forward contracts on commodities does make a difference to these results. The features of commodities are that there may be no chance to sell short, and storage is sometimes not possible if the commodity is perishable. This means the pricing relations have to be revised.

Returning to the basic strategies, it is possible to borrow money, buy the underlying asset, go short in a forward, hold the asset until the delivery date and then deliver and repay the loan. This must not be profitable.

Let U be present value of storage costs the strategy is not profitable if

$$F_0 \leq [S_0 + U]\, e^{rT}. \tag{15.24}$$

This relation puts an upper bound on the forward price. A lower bound cannot be applied without the possibility of short sales or of sales from stocks. If the good cannot be stored, then U can be thought of as the cost of actually producing the commodity.

15.8 Futures Compared to Forwards

In general, futures and forwards will not have the same price because of the daily settlement. This leads the two assets have different flows of payments.

When the risk-free interest rate is constant, then

$$\text{forward price} = \text{future price}. \tag{15.25}$$

This identity arises because with the constant interest rate the timing of the payments does not matter since they have the same present value.

Prices need not be the same when interest rates vary because of daily settlement. Consider a situation where the spot price, S, is positively correlated with the interest rate. With a long position, an increase in S earns a daily profit. Positive correlation ensures this is invested when r is high. Conversely, a decrease in S earns a loss which is covered when interest rates are low. This implies the future is more profitable than the forward.

Despite the observations, the difference in price may be small in practice.

15.9 Backwardation and Contango

The final issue to address is the relationship between the futures price and the expected spot price.

There are three possibilities that may hold.

1. Unbiased predictor.

In this case, the futures price is equal to the expected spot price at the delivery date of the contract. Hence

$$F_0 = E\left[S_T\right]. \tag{15.26}$$

2. Normal backwardation.

The argument for normal backwardation follows from assuming that

a. Hedgers will want to be short in futures,

b. Will have to offer a good deal to speculators,

Together these imply that

$$F_0 < E\left[S_T\right]. \tag{15.27}$$

3. Normal contango.

The argument for normal backwardation follows from assuming that

a. Hedgers will want to be long on average,

b. Must encourage speculators to be short,

Together these imply that

$$F_0 > E\left[S_T\right]. \tag{15.28}$$

The empirical evidence on this issue seems to suggest that generally $F_0 < E\left[S_T\right]$, so that normal backwardation holds.

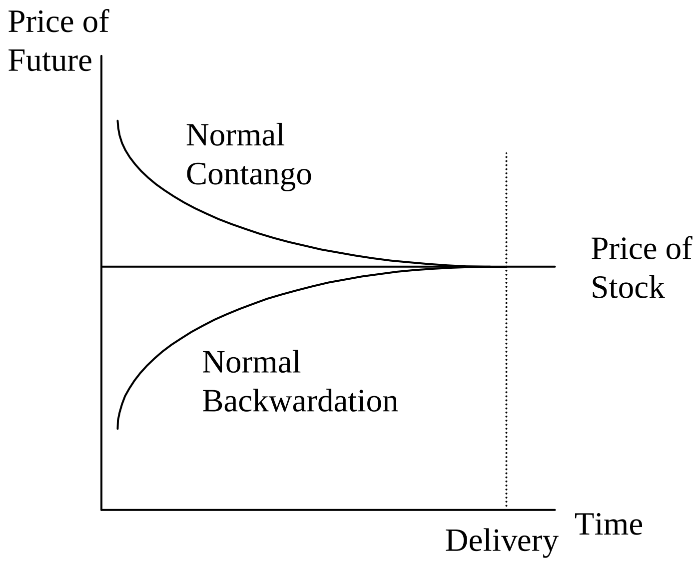


Figure 15.3: Backwardation and Contango

15.10 Conclusions

This chapter has introduced futures and forwards. The nature of the contracts has been described and the methods of valuation analyzed. A fair price has been determined by using both arbitrage arguments and the binomial model.

Exercise 107 *Explain normal backwardation and normal contango.*

Exercise 108 *A company has a \$36 million portfolio with a beta of 1.2. The S&P index future is currently standing at 900. Futures contracts on 250 times the index can be traded.*

(i) Derive the optimal hedge.

(ii) Determine the number of the contracts that should be traded and whether the position is long or short.

(ii) What if the beta is reduced to 0.9?

Exercise 109 *A bakery expects to need 100000 kilos of wheat in 3 months. The wheat futures contract is for the delivery of 30000 kilos of wheat.*

(i) How can the bakery use this for hedging?

(ii)From the bakery's point of view, what are the advantages and disadvantages of hedging?

Exercise 110 *Suppose that the standard deviation of quarterly changes in the price of a commodity is 0.8, the standard deviation of quarterly changes in a* *futures price on the commodity is 0.9, and the coefficient of correlation between the two changes is 0.75.*

(i) What is the optimal hedge ratio for a 3 month contract?

(ii) How is this interpreted?

Exercise 111 *A one-year-long forward contract on a non-dividend-paying stock is entered into when the stock price is £50 and the risk-free rate of interest is 10% per annum with continuous compounding.*

(i) What are the forward price and the initial value of the forward contract?

(ii) Six months later, the price of the stock is £55 and the risk-free interest rate is still 10%. What are the forward price and the value of the forward contract?

Chapter 16

Swaps

16.1 Introduction

In 1981 IBM and the World Bank undertook an exchange of fixed rate debt for floating rate debt. This exchange was the start of the interest rate swap industry. It is now estimated that the market is worth over $50 trillion per year. But it is difficult to provide a precise valuation of the size of the market because the market is not regulated and swaps are arranged between individual parties and not through exchanges.

The financial swaps we will consider are agreements to exchange one sequence of cash flows over a fixed period for another sequence of cash flows over the same period. This is precisely what IBM and the World Bank did.

The two sequences of cash flows are tied to either to a debt instrument or to a currency. This gives the two main types of swaps:

- Interest rate swaps
- Currency rate swaps

Why did swaps emerge? The first swaps were conducted in the late 1970s to avoid currency UK currency controls. These controls limited the value of currency that could be exchanged but this could easily be avoided by swapping rather than exchanging. These were followed by the IBM and World Bank swap in 1981. By 2001 it was estimated that $57 trillion in underlying value was outstanding in swap agreements.

The next section describes interest rate swaps and currency swaps. The use of swaps and the market for swaps are then described. The chapter then proceeds to the valuation of swaps.

16.2 Plain Vanilla Swaps

The basic form of interest rate swap, the *plain vanilla*, is now described.

The first step to do this is to introduce the LIBOR. This is the London Interbank Offered Rate – the rate of interest at which banks lend to each other. This rate is fundamental to valuing swaps since it acts as the basic "floating" rate of interest.

Definition 2 *British Bankers' Association (BBA) LIBOR is the BBA fixing of the London Inter-Bank Offered Rate. It is based on offered inter-bank deposit rates contributed in accordance with the Instructions to BBA LIBOR Contributor Banks. The BBA will fix BBA LIBOR and its decision shall be final. The BBA consults on the BBA LIBOR rate fixing process with the BBA LIBOR Steering Group. The BBA LIBOR Steering Group comprises leading market practitioners active in the inter-bank money markets in London. BBA LIBOR is fixed on behalf of the BBA by the Designated Distributor and the rates made available simultaneously via a number of different information providers. Contributor Panels shall comprise at least 8 Contributor Banks. Contributor Panels will broadly reflect the balance of activity in the inter-bank deposit market. Individual Contributor Banks are selected by the BBA's FX & Money Markets Advisory Panel after private nomination and discussions with the Steering Group, on the basis of reputation, scale of activity in the London market and perceived expertise in the currency concerned, and giving due consideration to credit standing. (http://www.bba.org.uk/bba/jsp/polopoly.jsp?d=225&a=1413)*

16.2.1 Interest Rate Swap

A swap requires two parties to participate. For the purpose of the discussion, call these party A and party B.

On one side of the swap, party A agrees to pay a sequence of fixed rate interest payments and to receive a sequence of floating rate payments. A is called the *pay-fixed* party.

On the other side of the swap, party B agrees to pay a sequence of floating rate payments and to receive a sequence of fixed rate payments. B is called the *receive-fixed* party

The *tenor* is the length of time the agreement lasts and the *notional principal* is the amount on which the interest payments are based. With a plain vanilla swap, interest is determined in advance and paid in arrears.

Example 181 *Consider a swap with a tenor of five years and two loans on which annual interest payments must be made. Let the notional principal for each loan be $1m. Party A agrees to pay a fixed rate of interest of 9% on the $1m. Party B receives this fixed rate, and pays the floating LIBOR to A.*

In principal, the swap involves loans of $1m being exchanged between the parties. That is, A has a floating interest rate commitment which is transfers to B and B has a fixed-rate commitment that it transfers to A. But in practice there is no need for these loans to exist and the principal can be purely nominal. In fact only the net payments, meaning the difference in interest payments, are made.

Table 16.1 illustrates the cash flows resulting from this swap agreement for a given path of the LIBOR. It must be emphasized that this path is not known when the swap agreement is made. The direction the LIBOR takes determines which party gains, and which party loses, from the swap. The parties will enter such an agreement if they find the cash flows suit their needs given the expectations of the path of the LIBOR.

Year, t	LIBOR$_t$	Floating Rate($B \to A$)	Fixed Rate($A \to B$)
0	8		
1	10	80,000	90,000
2	8	100,000	90,000
3	6	80,000	90,000
4	11	60,000	90,000
5	-	110,000	90,000

Table 16.1: Cash Flows for a Plain Vanilla Swap

16.2.2 Currency Swaps

A currency swap involves two parties exchanging currencies. It will occur when two parties each hold one currency but desire another. This could be for reasons of trade or because they aim to profit out of the swap based on expectations of exchange rate movements. The parties swap principal denominated in different currencies but which is of equivalent value given the initial exchange rate.

The interest rate on either principal sum may be fixed or floating. As an example, consider two parties C and D. Assume that C holds Euros but wants to have dollars. For instance, C may have to settle an account in dollars. In contrast, D holds dollars but wants to have Euros instead. The two parties can engage in a swap and trade the dollars for Euros. Unlike an interest rate swap, the principal is actually exchanged at the start of the swap. It is also exchanged again at the end of the swap to restore the currency to the original holder.

The fact that the interest rates can be fixed or floating on either currency means that there are four possible interest schemes:

- C pays a fixed rate on dollars received, D pays a fixed rate on Euros received
- C pays a floating rate on dollars received, D pays a fixed rate on Euros received
- C pays a fixed rate on dollars received, D pays a floating rate on Euros received
- C pays a floating rate on dollars received, D pays a floating rate on Euros received

The predominant form of contract is the second. If party D is a US firm, then with a *plain vanilla currency swap* the US firm will pay a fixed rate on the currency it receives.

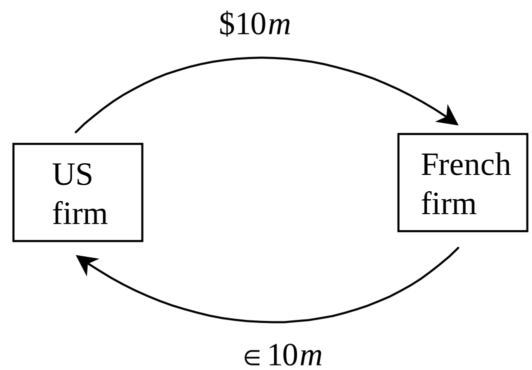


Figure 16.1: Currency Swap

To show how a currency swap functions, consider a swap of type 1 which involves exchanging fixed-for-fixed. The cash flows that occur with this swap are:

- The initial swap of currency at initiation
- The periodic interest payments
- The swap of principal at termination

A currency swap involves interest payments which are made in the currency received. Consequently, since the two payments are in different currencies, there is no netting of the interest payments.

Example 182 *Consider a US firm that holds dollars but wants euros and a French firm that holds Euros but wants dollars. Both parties agree to pay fixed interest. Assume that:*
a. The spot exchange rate is $1 = €1. The spot rate is the rate for immediate exchange of currency.
b. The US interest rate is 10%
c. The French interest rate is 8%
d. The tenor of the swap is 6 years
e. Interest is paid annually
f. The principal swapped is $10m for €10m.

It should be noted that given the spot exchange rate, the principal exchanged is of equal value. This implies the fact that the swap must always be of equal value at the initial spot exchange rate. Figure 16.1 displays the exchange of principal at the start of the swap.

The cash flows during the period of the swap are illustrated in Table 16.2. This shows that the interest payments are made in the currency received. Since the swap is fixed-for-fixed, the interest payments remain constant.

	To US	From US	To French	From French
0	$\in 10m$	\$10m	\$10m	$\in 10m$
1	\$1m	$\in 0.8m$	$\in 0.8m$	\$1m
2	\$1m	$\in 0.8m$	$\in 0.8m$	\$1m
3	\$1m	$\in 0.8m$	$\in 0.8m$	\$1m
4	\$1m	$\in 0.8m$	$\in 0.8m$	\$1m
5	\$1m	$\in 0.8m$	$\in 0.8m$	\$1m
6	\$1m	$\in 0.8m$	$\in 0.8m$	\$1m

Table 16.2: Cash Flow for Fixed-for-Fixed

Given these payments, it is natural to ask which flow is best. The answer to this question depends on (i) the needs of the two firms for currency, and (ii) the course of exchange rates over the lifetime of the swap. Because the interest and principal have to be repaid in a currency different to the one that was initially held, entering a swap agreement opens the parties up to exchange rate risk.

Example 183 *Consider a swap between a US firm and a Japanese firm. The Japanese firm pays a floating rate on dollars received and the US firm pays a fixed rate on the Yen received. Assume that:*

a. The spot exchange rate be \$1 = Y120.

b. The principal is \$10m when denominated in dollars and Y1200m when denominated in Yen.

c. The tenor of the swap is 4 years.

d. The Japanese 4-year fixed interest rate is 7%. This is the interest rate paid on the Yen received by the US firm.

e. The rate on the dollar is the LIBOR, which is 5% at the initiation of the swap.

The cash flows during the swap are determined by the path of the LIBOR. Table 16.3 displays the flows for one particular path of the LIBOR. In this table, the LIBOR rises over time so the interest payments received by the US firm increase over time. If the exchange rate were constant, this would be advantageous for the US firm. However, as will be seen later, the exchange rate is related to the interest rate and this needs to be taken into account before this claim can be established.

Time	LIBOR	Japanese in	Japanese out	US in	US out
0	5%	\$10m	Y1200m	Y1200m	\$10m
1	6%	Y84m	\$0.5m	\$0.5m	Y84m
2	7%	Y84m	\$0.6m	\$0.6m	Y84m
3	10%	Y84m	\$0.7m	\$0.7m	Y84m
4		Y1284m	\$1.1m	\$1.1m	Y1284m

Table 16.3: Cash Flow on Fixed-for-Floating

16.3 Why Use Swaps?

There are three major reasons why swaps may be used. These are now considered in turn.

16.3.1 Market Inefficiency

A first reason for using swaps is to overcome market inefficiency. For example, it could be the case that firms located in a country are able to borrow at a lower rate in that country than firms located abroad. This creates a position in which firms have a comparative advantage in borrowing in their country's currency. Given such a position of comparative advantage, it is possible for two parties to find a mutually advantageous trade.

Such a trade is illustrated in Table 16.4 where the US firm can borrow dollars at 9% but the UK firm must pay 10% to borrow dollars. The opposite position holds for borrowing in the UK.

	US \$ rate	UK £ rate
US firm	9%	8%
UK firm	10%	7%

Table 16.4: Interest Rates

Assume that the UK firm wants dollars and the US firm wants Sterling. If they were to borrow directly at the rates in the table, the US firm would pay a rate of interest of 8% on its sterling and the UK firm a rate of 10% on its dollars.

If the firms were to borrow in their own currency and then swap, this would reduce the rate faced by the US firm to 7% and that faced by the UK to 9%. This swap is illustrated in Figure 16.2. The exploitation of the comparative advantage is beneficial to both parties.

The existence of the comparative advantage depends on there being a market inefficiency that gives each firm an advantage when borrowing in its home market. If the market were efficient, there would be a single ranking of the riskiness of the firms and this would be reflected in the rates of interest they pay in both countries. The internalization of financial markets makes it unlikely that there will be significant inefficiencies to be exploited in this way.

16.3.2 Management of Financial Risk

Swaps can be used to manage financial risk. This is clearest when assets and liabilities are mismatched.

The US Savings and Loans provide a good example of the possibility of risk management using swaps. These institutions receive deposits from savers and use the funds to provide loans for property.

The Savings and Loans pay floating rate interest on deposits but they receive fixed rate interest on the loans they grant. Since the loans are for property they are generally very long term.

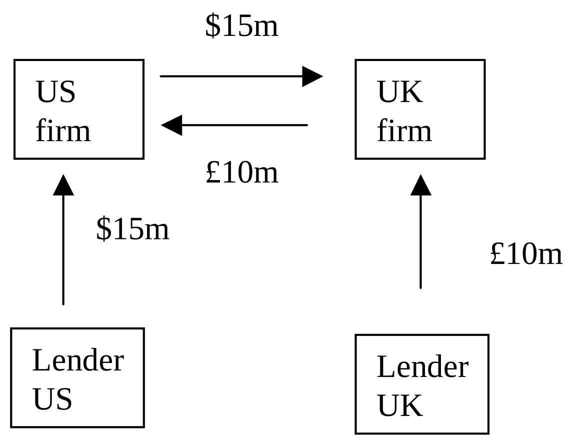


Figure 16.2: Exploiting Comparative Advantage

This places the Savings and Loans in a position where they are exposed to risk if the floating interest rate rises. Such a rise would create an increase in their payments to depositors but would not be accompanied by any increase from the long-term loans. Precisely this position was responsible, at least in part, for the collapse of a number of these institutions in the 1980s.

Example 184 *The Savings and Loan crisis of the 1980s was a wave of savings and loan failures in the USA, caused by mismanagement, rising interest rates, failed speculation and, in some cases, fraud. U.S. taxpayers took the brunt of the ultimate cost, which totaled around US$600 billion. Many banks, but particularly savings and loan institutions, were experiencing an outflow of low rate deposits, as depositors moved their money to the new high interest money market funds. At the same time, the institutions had much of their money tied up in long term mortgages which, with interest rates rising, were worth far less than face value. Early in the Reagan administration, savings and loan institutions ("S&Ls") were deregulated (see the Garn - St Germain Depository Institutions Act of 1982), putting them on an equal footing with commercial banks. S&Ls (thrifts) could now pay higher market rates for deposits, borrow money from the Federal Reserve, make commercial loans, and issue credit cards. (http://en.wikipedia.org/wiki/Savings_ and_ Loan_ scandal)*

A solution to the risk problem faced by the Savings and Loans would have been to swap the fixed interest rate loans for floating interest rate loans. By doing this, they could have ensured that any increase in interest rates is met by an increase in expenditure and receipts.

It is clear that there are other possible responses for the Savings and Loans to secure their position. For instance they could issue bonds with a fixed coupon and buy floating rate notes. This would then balance their portfolio as a whole.

The reason why such a trade may not be used is because of legislation which limits the financial activities that can be undertaken.

16.3.3 Speculation

Expressed in the simplest terms, a swap is a no more than a bet on the direction of interest rate and/or exchange rate movements. If the movement is in the right direction, a profit can be earned. Swaps can therefore be used for purely speculative reasons.

16.4 The Swap Market

This section discusses the major features of the swap market and the participants in that market.

16.4.1 Features

The major features of the swap market are the following:

a. There is no publicly observable marketplace. Swaps are transactions that take place either between individuals directly, between individuals with the intermediation of a broker, or with a swap dealer. Brokers and dealers are discussed further below.

b. There is limited government regulation. Since there is no marketplace it is difficult for any government to provide regulation. There has been some recent discussion of regulation.

Example 185 *America's continued financial leadership in the new economy is at stake as Congress sets out to modernize the Commodity Exchange Act, the law that covers futures and derivatives trading. The revised law is supposed to liberalize the derivatives market, setting important legal terms that distinguish traditional commodity futures from over-the-counter derivatives, or swaps. The future of the U.S. swaps market depends on whether Congress can keep it free of entangling regulations and legal uncertainty. Derivatives are an essential tool of risk management for American businesses. They are the lubricants that let financial markets allocate capital more efficiently. Foreign exchange swaps, for example, diminish the risks associated with fluctuating currencies. Rate swaps smooth out the effects of interest-rate fluctuations by converting long-term, fixed-rate debt into short-term, variable-rate debt. OTC derivatives make businesses more competitive by lowering their cost of capital. To be effective, the enforceability and legal status of swaps must be firmly established. Banks and other financial institutions have worried for years that the Commodity Futures Trading Commission might begin applying futures regulations to swaps. That would be disastrous, since futures contracts are legally enforceable only if they are traded on a listed exchange, such as the Chicago Mercantile Exchange. Off-exchange swaps are privately negotiated, custom-tailored contracts. Trillions of dollars in interest-rate and currency-swap contracts would be undermined if they were* *suddenly regulated like futures. Banks furnishing swaps to large institutional and corporate clients are poised to extend the benefits of these risk management tools to their small business and retail customers. But they are wary of the CFTC, which in 1998 considered regulating swaps. The legal uncertainty this created was unsettling to the financial markets, which don't consider the CFTC technically competent to regulate complex swap transactions. Unwarranted bureaucratic restrictions would reduce the technical precision of swaps and increase their cost. House bill H.R. 4541, the Commodity Futures Modernization Act, is supposed to rationalize the regulatory environment and provide legal certainty. But this effort is fragmented because of the competing jurisdictions of regulatory agencies and congressional committees. An amendment recently offered by House Banking Committee Chairman Jim Leach, R-Iowa, goes the furthest in liberalizing OTC swaps, but still leaves room for regulatory meddling. Though the CFTC couldn't regulate them, the Treasury Department or Federal Reserve could. Other versions of H.R. 4541 set up a convoluted series of exemptions to insulate most swaps from CFTC regulation, but don't exempt the entire universe of swaps. Individual investors worth less than $5 million to $10 million in assets will likely face regulatory hurdles. Ostensibly these restrictions are meant to protect retail investors from fraud. However, as Harvard University law professor Hal Scott testified to the House Banking Committee, the true purpose might be "to fence off exchange-traded derivatives markets from competition with OTC derivatives markets for retail investors." Swap contracts completed over electronic trading facilities are potentially vulnerable under the bill. Specifically, derivative transactions resulting from "automated trade matching algorithms" are exposed to additional regulation. This language could inhibit the new economy innovators that match trades electronically using highly specialized software. The big commodity exchanges would benefit from rules that hinder off-exchange innovators. But the added red tape will only delay the inevitable. If regulatory barriers are set up to protect the futures industry from electronic competition, the innovators will simply move offshore. If Congress wishes to liberalize swaps, it should do so by defining commodity futures narrowly and prohibiting any regulation of OTC derivatives outside the definition. Over-the-counter swaps should be completely exempt from antiquated exchange rules that were designed for the old economy. Rather than leaving any OTC derivatives in regulatory limbo, Congress should confer ironclad legal certainty upon all kinds of swaps. (Swap New For Old: Congress Shouldn't Impose Tired Rules On OTC Derivatives by James M. Sheehan, August 9, 2000, Investor's Business Daily)*

c. Contracts cannot be terminated early. The nature of a swap deal is that it is a commitment that must be seen through to the end. Once it is agreed it is not possible to withdraw from the deal.

d. No guarantees of credit worthiness. With futures there is an exchange which manages the contracts to avoid any possibility of default by ensuring margin is held and limiting daily movements of prices. The fact that there is no marketplace for swaps implies that there is no similar institution in the swap market.

Example 186 *London Borough of Hammersmith and Fulham: A local government in the United Kingdom that was extremely active in sterling swaps between 1986 and 1989. Swap volume was very large relative to underlying debt, suggesting large scale speculation by the borough council. The speculation was unsuccessful and a local auditor ruled that the transactions were ultra vires-beyond the powers of the council. The House of Lords sitting as the High Court ultimately upheld the auditor's ruling. The "legal" risk of some risk management contracts was established at considerable cost to the London financial community. (http://riskinstitute.ch/00011654.htm)*

16.4.2 Dealers and Brokers

For anyone wishing to conduct a swap there is the problem of finding a counterparty. For other derivatives, such as options and futures, this is less of a problem since there are organized exchanges to assist with transactions.

In the early days of the swap market counterparties to a swap were originally found via a broker. The market has developed so that swaps are now generally conducted through dealers. This has increased the efficiency of the swap market.

Swap Broker

A swap broker acts as an intermediary in the market. Their role is to match swap parties who have complementary needs.

A broker maintains a list of clients who are interested in entering into swap deals and tries to match the needs of the clients.

But because it is necessary for a broker to find matching clients before any trade can take place, the organization of a market through brokers does not make for a very efficient market.

Swap Dealer

A swap dealer acts as a counter-party to a swap. They can be on either side of the deal. The profit of a swap dealer is obtained by charging a spread between the two sides of the deal.

The dealer accumulates a *swap book*. The book is constructed with the aim: of balancing trades to limit risk.

The risks facing a swap dealer are the following:

1. Default risk

The party on the other side of a swap may default.

2. Basis risk

The basis risk arises from movements in interest rates.

3. Mismatch risk

Mismatch risk arises from the two sides of the dealers swap book not being balanced.

16.5 The Valuation of Swaps

The process of valuation relates to answering two related questions. How is a swap correctly priced? How can the deal be fair for both parties?

As an example, consider a plain vanilla interest rate swap. The party on one side of this swap will pay the floating LIBOR rate, while the party on the other side pays a fixed rate of interest. The only variable in this transaction that can be adjusted to make the deal fair for both parties is the fixed rate. By making this higher, the receive-fixed party benefits. Make it lower and the pay-fixed party benefits.

The fundamental issue is to determine what fixed rate should be used to make the deal fair. Here fair means that both parties see the swap as equally advantageous at the time at which it is agreed.

Before proceeding to determine the fixed rate, it is worth looking at how swaps are related to bond portfolios. The reasoning is the same as that applied to options and forwards: the swap is constructed so that there are no arbitrage opportunities. Both of the earlier derivatives were priced by constructing a replicating portfolio that gave the same payoffs as the derivative. Applying the arbitrage argument then means the price of the derivative must be the same as the cost of the replicating portfolio.

The same basic logic can be applied to swaps where bonds can be used to replicate the position of a party who has entered a swap deal.

16.5.1 Replication

Definition: a floating rate note is a bond that pays a floating rate of interest (LIBOR for this analysis)

1. Interest rate swaps

a. Plain vanilla receive-fixed

This is equivalent to
-a long position in a bond
-a short position in a floating rate note

Example 1. A 6% corporate bond with annual coupon maturity 4 years, market value of $40m trading at par

2. A floating rate note, $40m principal, pays LIBOR annually, 4 year maturity

The cash flows are shown in Figure 16.3.

These flows match those for a swap with notional principal of $40m and a fixed rate of 6%.

b. Plain Vanilla Pay-Fixed

The swap is equivalent to:
-issue bond (go short) a fixed-coupon bond
-but (go long) a floating rate note

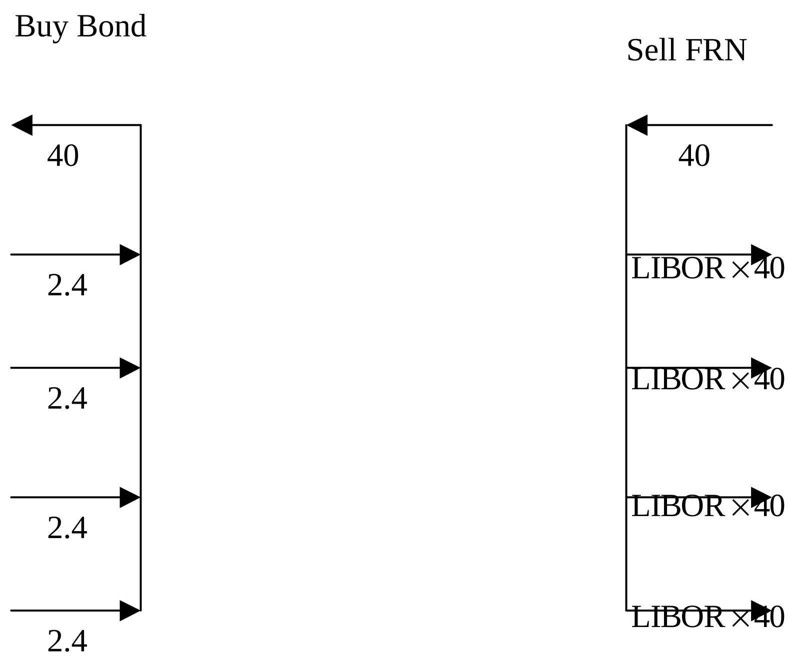
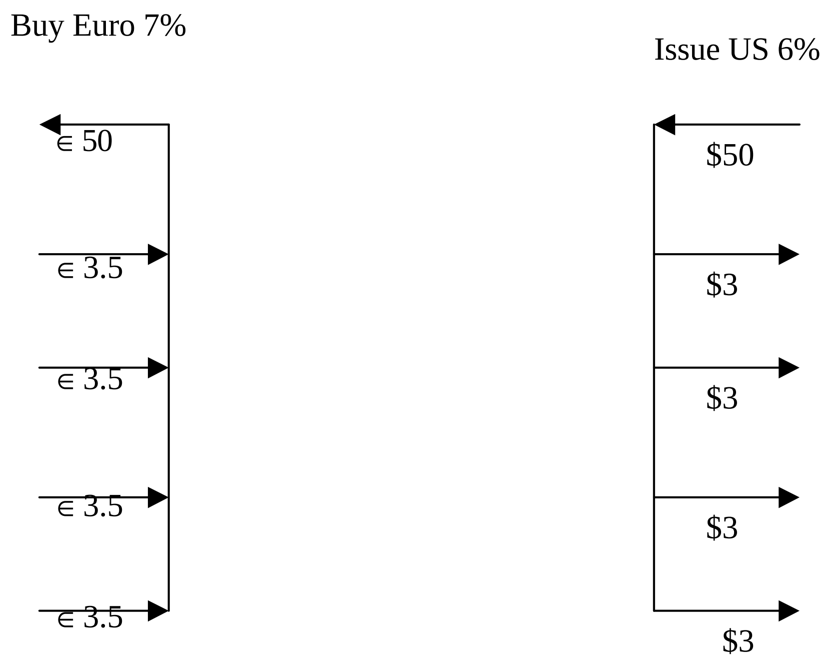


Figure 16.3: Cash Flows



2. Currency Swaps

a. Fixed-for-Fixed Currency Swap
-buy a bond in one currency
-issue bond denominated in another
b. Plain Vanilla Currency Swap
-one bond fixed coupon
-one floating rate note

16.5.2 Implications

1. Motive for swaps?
Economize on cost of these bond portfolios
2. Pricing of swaps?
Since they can be replicated by bonds, must be related to interest rates on bonds

16.6 Interest Rate Swap Pricing

The essential item to be determined in pricing an interest rate swap is to set the fixed interest rate so, given that the other party pays LIBOR, the swap is fair.

To see how the argument functions, consider a plain vanilla interest rate swap. The receive-fixed party pays LIBOR. The fixed rate has to be set so that there are no arbitrage opportunities. Define the SFR as the *Swap Fixed Rate.* This is the fixed rate that will be constructed to make the swap fair.

For there to be no arbitrage, the two flows of payments over the life of the swap must have the same present value. This present value has to be computed using the rates of interest observed in the market. Fundamental to this process is the term structure and the implied forward rates. The term structure is the set of spot interest rates for spot loans of different lengths. These spot rates imply the forward rates. This was covered in Chapter 12.

Consider a swap with notional principal of $1m and a tenor of 4 years. The floating interest rate in each year is predicted by the forward rate. Note that these rates are all observed at the time the swap is organized and contracts can be made to borrow and lend at these rates of interest. They need not, and almost certainly will not, be the rates that actually hold when the future periods are reached but they are the best predictor at the start of the swap.

Year	Floating Rate	Fixed Rate
1	$f_{0,1}$	SFR
2	$f_{1,2}$	SFR
3	$f_{2,3}$	SFR
4	$f_{3,4}$	SFR

Table 16.5: Interest Rates

Using the interest rates in Table 16.5, the present value of the cash flows must be equal. Given that the value of the notional principal is $1m, the present value of the series of floating interest payments is

$$PV(floating) = \frac{f_{0,1}}{1+s_1} + \frac{f_{1,2}}{[1+s_2]^2} + \frac{f_{2,3}}{[1+s_3]^3} + \frac{f_{3,4}}{[1+s_4]^4}. \tag{16.1}$$

The present value of the fixed interest payments is

$$PV(fixed) = \frac{SFR}{1+s_1} + \frac{SFR}{[1+s_2]^2} + \frac{SFR}{[1+s_3]^3} + \frac{SFR}{[1+s_4]^4}. \tag{16.2}$$

Equating these two present values and solving, the SFR can be found to be

$$SFR = \frac{\sum_{n=0}^{3} \frac{f_{n,n+1}}{[1+s_{n+1}]^{n+1}}}{\sum_{m=0}^{4} \frac{1}{[1+s_m]^m}}. \tag{16.3}$$

This is the swap fixed rate that leads to no arbitrage being possible since it equates the present values.

Note further that the relation between spot rates and forward rates makes it possible to translate between the two. In particular,

$$1 + s_1 = 1 + f_{0,1}, \tag{16.4}$$

$$[1 + s_2]^2 = [1 + f_{0,1}][1 + f_{1,2}], \tag{16.5}$$

$$[1 + s_3]^3 = [1 + f_{0,1}][1 + f_{1,2}][1 + f_{2,3}], \tag{16.6}$$

$$[1 + s_4]^4 = [1 + f_{0,1}][1 + f_{1,2}][1 + f_{2,3}][1 + f_{3,4}], \tag{16.7}$$

Using these relations, SFR can be expressed either:

1. In terms of spot rates

or

2. In terms of forward rates.

Example 187 *Let the spot rates be* $s_1 = 4\%, s_2 = 5\%, s_3 = 6\%, s_4 = 7\%$. *Then* $f_{0,1} = 4\%,\ f_{1,2} = 6\%,\ f_{2,3} = 8\%,\ f_{3,4} = 10\%$. *So*

$$SFR = \frac{\frac{0.04}{1.04} + \frac{0.06}{[1.05]^2} + \frac{0.08}{[1.06]^3} + \frac{0.1}{[1.07]^4}}{\frac{1}{1.04} + \frac{1}{[1.05]^2} + \frac{1}{[1.06]^3} + \frac{1}{[1.07]^4}} = 0.068 \textit{ (6.8\%)}$$

In general, if interest is paid at intervals of length τ and the tenor of the swap is $N\tau$, then the formula for the swap fixed rate can be generalized to

$$SFR = \frac{\sum_{n=1}^{N} \frac{f_{[n-1]\tau,n\tau}}{z_{0,n\tau}}}{\sum_{m=1}^{N} \frac{1}{z_{0,m\tau}}}, \tag{16.8}$$

where $z_{0,n\tau}$ is the discount factor between time 0 and time $n\tau$.

These results determine what the fixed rate should be in the swap to match the floating LIBOR.

16.7 Currency Swap

With a currency swap there is the additional feature of changes in the exchange rate. This requires an extension to the analysis. The extension has to relate the swap fixed rates in the two countries to the term structure in both countries and the exchange rates.

16.7.1 Interest Rate Parity

Consider two countries A and B. The information that is available at the initiation of the swap consists of:

1. The term structure in A
2. The term structure in B
3. The rates for foreign exchange between the currencies of the two countries.

Under (3) we observe both the spot exchange rates and the forward exchange rates. Forward exchange rates give the rate now for an agreed currency exchange at a fixed date in the future.

The notation is to use ${}_{AB}e_{0,0}$ to denote the value at time 0 for currency A in terms of currency B for delivery at 0. This is the spot exchange rate. For instance, if £1(currency A) = \$1.5 (currency B) then ${}_{AB}e_{0,0} = 1.5$.

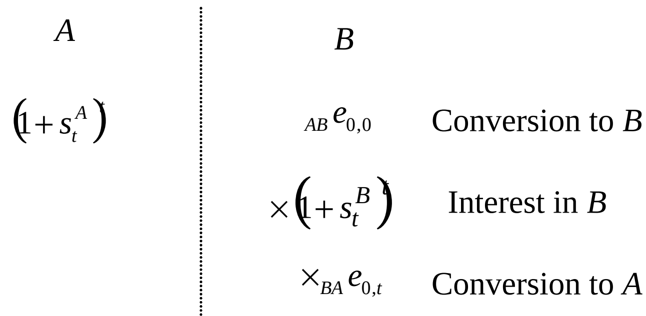


Figure 16.4: Interest Rate Parity

Similarly, the notation $_{AB}e_{0,t}$ denotes the value contracts made at time 0 for currency A in terms of currency B for delivery of the currency at time t. This is a forward exchange rate.

These exchange rates do not stand alone but are related via the spot rates of interest. This is a consequence of the fact that transactions can be undertaken to trade the currencies at spot and forward rates.

Consider the following two investment strategies:

- Invest 1m in country A for t years
- Convert 1m to currency of country B and invest for t years and enter forward to convert back

The basis of this strategy is that all the interest rates and exchange rates are known at time 0 so the cash flows are certain. The fact that everything is certain implies that the payoffs of the two strategies must be the same. If they were not, then arbitrage would take place. The two strategies are shown in Figure 16.4.

To eliminate the possibility of arbitrage it must be the case that

$$\left(1+s_t^A\right)^t =_{AB} e_{0,0}\left(1+s_t^B\right)^t_{BA} e_{0,t}, \tag{16.9}$$

so that given the spot rates it is possible to calculate the currency forward rates. These currency forward rates can then be used these to obtain the present value of a swap deal at the initiation of the swap.

The claim made here is that interest rate parity connects SFR^A to SFR^B. If it did not then there would be arbitrage between the currencies of the two countries. Therefore it is possible to use the SFR in each country as the fixed rate in a currency swap.

16.7.2 Fixed-for-Fixed

Consider a fixed-for-fixed swap involving an exchange of dollars for a “foreign” currency.

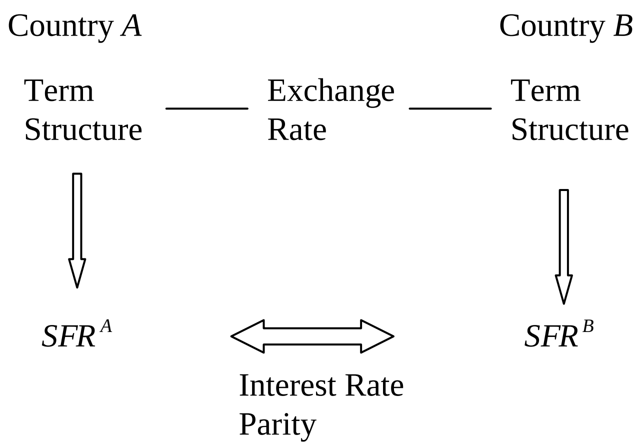


Figure 16.5: Interest Rates and Exchange Rates

Let A be the party that receives dollars and pays a fixed rate on these dollars.

Let B be the party that receives the "foreign" currency and also pays a fixed rate on this foreign currency.

To determine the fair value of the swap, the issue is to determine what fixed rates should be used.

The answer that will be demonstrated is that:

Party A: pays dollar SFR - the SFR on a corresponding dollar plain vanilla interest rate swap

Party B: pays "foreign" SFR - the SFR on a corresponding "foreign" plain vanilla interest rate swap

Doing this ensures the present value of expected cash flows for A and B are zero.

Two demonstrations of this are given. The first is taken from the text by Kolb and involves adopting a set of numbers and evaluating an example. The second demonstration shows the result algebraically for a swap with a very short tenor.

Demonstration A

Consider a \$ for Dm swap. Assume that the spot rate for current exchange is \$1 = DM2.5. Let the principal on the swap be \$100m. This is equal to DM250m at the initial spot rate. The tenor of the swap is 5 years.

The first step in constructing the correct values of the SFRs is to use the term structure in each country to generate the implied path of the exchange rate. The interest rate parity argument in (16.9) gives the relationship between the spot rates and exchange rates as

$$_{\$DM}e_{0,t} = \frac{\left(1 + s_t^{DM}\right)^t}{_{DM\$}e_{0,0}\left(1 + s_t^{\$}\right)^t}. \tag{16.10}$$

This formula determines the forward exchange rates for the two currencies. It should also be noted that by definition, the two exchanges rates are related by

$$_{DM\$}e_{0,0} = \frac{1}{_{\$DM}e_{0,0}}. \tag{16.11}$$

To allow a numerical demonstration, Table 16.6 assumes values for the $ and DM term structures. Combining these with interest rate parity, the implied path of the exchange rate can be derived. This is given in the final column.

Year	$s_t^{\$}$	$\left(1+s_t^{\$}\right)^t$	s_t^{DM}	$\left(1+s_t^{DM}\right)^t$	$_{\$DM}e_{0,t}$
0	-	-	-	-	2.50
1	0.08	1.08	0.05	1.05	2.430
2	0.085	1.177	0.052	1.106	2.349
3	0.088	1.289	0.054	1.171	2.71
4	0.091	1.421	0.055	1.240	2.181
5	0.093	1.567	0.056	1.315	2.097

Table 16.6: Term Structures and Exchange Rate

The second step is to use the term structure to calculate the implied set of forward rates. These are shown in Table 16.7.

	$	DM
$f_{0,1} = s_1$	0.08	0.5
$f_{1,2} = \frac{(1+s_2)^2}{(1+s_1)} - 1$	0.089	0.053
$f_{2,3} = \frac{(1+s_3)^3}{(1+s_2)^2} - 1$	0.095	0.058
$f_{3,4} = \frac{(1+s_4)^4}{(1+s_3)^3} - 1$	0.102	0.059
$f_{4,5} = \frac{(1+s_5)^5}{(1+s_4)^4} - 1$	0.103	0.0605

Table 16.7: Forward Rates

The third step is to use these forward rates to generate the swap fixed rates through the formula

$$SFR = \frac{\sum_{t=1}^{5} \frac{f_{t-1,t}}{(1+s_t)^t}}{\sum_{t=1}^{5} \frac{1}{(1+s_t)^t}}. \tag{16.12}$$

Using the values in Table 16.7, the two swap fixed rates are

$$\begin{aligned} SFR^{\$} &= \frac{0.074 + 0.076 + 0.074 + 0.072 + 0.066}{0.926 + 0.850 + 0.776 + 0.704 + 0.638} \\ &= 0.929, \end{aligned} \tag{16.13}$$

and

$$SFR^{DM} = 0.056. \tag{16.14}$$

The $SFR^{\$}$ in (16.13) is the value that would be used in a \$ interest rate swap and the SFR^{DM} in (16.14) is the rate that would be used in a DM interest rate swap. These are the values that are consistent with the elimination of arbitrage possibilities and make the swap fair for both parties.

The fact that these are the correct SFRs can be shown by using these values to determine the expected cash flows during the life of the swap It should be noted that these are the flows expected given the observed term structures. If future interest rates are not as implied by the term structure, then the actual cash flows will be different.

TABLE OF APPLICATION

The conclusion derived from observing the figures in this table is that these SFR values do give a fair price so the swap is of fair value for both parties. The initial present value of the swap, evaluated using interest rate parity to determine the exchange rates, is zero for both parties.

Demonstration B

The second demonstration that the SFR is the correct rate to use undertakes the calculations using the general definitions of the variables.

Consider a swap of DM for \$ with a two-year tenor. Table 16.8 states the cash flows for the two parties involved with the swap per \$ of principal.

Year	DM cash flow	\$ cash flow	DM value of \$
0	$-_{\$DM}e_{0,0}$	1	$_{\$DM}e_{0,0}$
1	$_{\$DM}e_{0,0}SFR^{DM}$	$-SFR^{\$}$	$-_{\$DM}e_{0,1}SFR^{\$}$
2	$_{\$DM}e_{0,0}\left(1+SFR^{DM}\right)$	$-\left(1+SFR^{\$}\right)$	$-_{\$DM}e_{0,2}\left(1+SFR^{\$}\right)$

Table 16.8: Cash Flows

The next table presents the net DM cash flows.

Year	Net DM cash flow	Discount on DM
0	$_{\$DM}e_{0,0} -_{\$DM} e_{0,0} = 0$	1
1	$_{\$DM}e_{0,0}SFR^{DM} -_{\$DM} e_{0,1}SFR^{\$}$	$\frac{1}{\left(1+s_1^{DM}\right)}$
2	$_{\$DM}e_{0,0}\left(1+SFR^{DM}\right) -_{\$DM} e_{0,2}\left(1+SFR^{\$}\right)$	$\frac{1}{\left(1+s_2^{DM}\right)^2}$

Table 16.9: Net Cash Flows

The present value of the DM cash flow is

$$\begin{aligned} PV = {} & \frac{1}{\left(1+s_1^{DM}\right)}\left[_{\$DM}e_{0,0}SFR^{DM} -_{\$DM} e_{0,1}SFR^{\$}\right] \\ & + \frac{1}{\left(1+s_2^{DM}\right)^2}\left[_{\$DM}e_{0,0}\left(1+SFR^{DM}\right) -_{\$DM} e_{0,2}\left(1+SFR^{\$}\right)\right]. \end{aligned} \tag{16.15}$$

By definition, the forward exchange rates are

$$_{\$DM}e_{0,1} = \frac{\left(1+s_1^{DM}\right)}{_{DM\$}e_{0,0}\left(1+s_1^{\$}\right)}, \tag{16.16}$$

$$_{\$DM}e_{0,2} = \frac{\left(1+s_2^{DM}\right)^2}{_{DM\$}e_{0,0}\left(1+s_2^{\$}\right)^2}, \tag{16.17}$$

and

$$_{DM\$}e_{0,0} = \frac{1}{_{\$DM}e_{0,0}}. \tag{16.18}$$

Using these exchange rates, the present value is

$$\begin{aligned} PV = {} & \frac{1}{\left(1+s_1^{DM}\right)}\left[_{\$DM}e_{0,0}SFR^{DM} -_{\$DM} e_{0,0}\frac{\left(1+s_1^{DM}\right)}{\left(1+s_1^{\$}\right)}SFR^{\$}\right] \\ & + \frac{1}{\left(1+s_2^{DM}\right)^2}\left[_{\$DM}e_{0,0}\left(1+SFR^{DM}\right) -_{DM\$} e_{0,0}\frac{\left(1+s_2^{DM}\right)^2}{\left(1+s_2^{\$}\right)^2}\left(1+SFR^{\$}\right)\right], \end{aligned} \tag{16.19}$$

or, simplifying this expression,

$$PV =_{\$DM} e_{0,0}\left[\left(\frac{SFR^{DM}}{\left(1+s_1^{DM}\right)} + \frac{\left(1+SFR^{DM}\right)}{\left(1+s_2^{DM}\right)^2}\right) - \left(\frac{SFR^{\$}}{\left(1+s_1^{\$}\right)} + \frac{\left(1+SFR^{\$}\right)}{\left(1+s_2^{\$}\right)^2}\right)\right]. \tag{16.20}$$

The swap fixed rate is defined by

$$\begin{aligned} SFR &= \frac{\frac{f_{0,1}}{1+s_1} + \frac{f_{1,2}}{(1+s_2)^2}}{\frac{1}{1+s_1} + \frac{1}{(1+s_2)^2}} \\ &= \frac{\frac{s_1}{1+s_1} + \frac{\frac{(1+s_2)^2}{1+s_1}-1}{(1+s_2)^2}}{\frac{1}{1+s_1} + \frac{1}{(1+s_2)^2}} \\ &= \frac{(1+s_2)^2(1+s_1)-(1+s_1)}{(1+s_1)+(1+s_2)^2}. \end{aligned} \tag{16.21}$$

The SFR can be substituted into the definition for present value (16.20) to give

$$\begin{aligned} PV &= {}_{\$DM}e_{0,0}\left[\begin{array}{c} \frac{SFR^{DM}\left[\left(1+s_1^{DM}\right)+\left(1+s_2^{DM}\right)^2\right]+\left(1+s_1^{DM}\right)}{\left(1+s_1^{DM}\right)\left(1+s_2^{DM}\right)^2} \\ -\frac{SFR^{\$}\left[\left(1+s_1^{\$}\right)+\left(1+s_2^{\$}\right)^2\right]+\left(1+s_1^{\$}\right)}{\left(1+s_1^{\$}\right)\left(1+s_2^{\$}\right)^2} \end{array}\right] \\ &= {}_{\$DM}e_{0,0}\left[\begin{array}{c} \frac{\left(1+s_2^{DM}\right)^2\left(1+s_1^{DM}\right)-\left(1+s_1^{DM}\right)+\left(1+s_1^{DM}\right)}{\left(1+s_1^{DM}\right)\left(1+s_2^{DM}\right)^2} \\ -\frac{\left(1+s_2^{\$}\right)^2\left(1+s_1^{\$}\right)-\left(1+s_1^{\$}\right)+\left(1+s_1^{\$}\right)}{\left(1+s_1^{\$}\right)\left(1+s_2^{\$}\right)^2} \end{array}\right] \\ &= {}_{\$DM}e_{0,0}\left[1-1\right] \\ &= 0. \end{aligned} \tag{16.22}$$

This completes the demonstration that the present value of the swap is zero.

16.7.3 Pricing Summary

This use of the SFR in a fixed-for-fixed swap provides the insight necessary to understand the interest rates used in other swaps.

A convenient summary of the results is the following:

1. Fixed-for-Fixed

Both parties pay the SFR for the currency received.

2. Floating-for-Fixed

The pay-floating party pays LIBOR, and the pay-fixed pays SFR (The LIBOR rate is that on the currency received).

3. Fixed-for-Floating

The pay-fixed party pays SFR, and the pay-floating pays LIBOR (The LIBOR rate is that on the currency received).

4. Floating-for-Floating

Both parties pay the LIBOR on the currency received.

16.8 Conclusions

This chapter has introduced swaps and the swap markets. It has also been shown how these swaps can be priced by setting the swap fixed rate to give the swap the same present value for the two parties on either side of the swap.

Exercise 112 *Assume that a US and UK firm engage in a currency swap. Let the spot exchange rate at the time of the swap be £1 = $1.60, the LIBOR rate be 5% and the fixed UK £rate be 6%. If the principal is £10m, chart the cash flows for the two parties when the tenor is 5 years.*

Exercise 113 *Consider a swap dealer with the following swap book.*

Swap	*Notional Principal (£ million)*	*Tenor (Years)*	*Fixed Rate (%)*	*Dealer's Position*
A	*10*	*4*	*7*	*Receive-Fixed*
B	*35*	*3*	*6.5*	*Pay-Fixed*
C	*20*	*5*	*7.25*	*Pay-Fixed*
D	*40*	*4*	*7.5*	*Receive-Fixed*
E	*15*	*1*	*6.75*	*Receive-Fixed*

If the applicable LIBOR rate is currently 5% but rises 1% per year, determine the yearly cash flow of the dealer if no further deals are made.

What should the dealer do to reduce their risk? [6 marks]

Exercise 114 *Consider the following term structures:*

Year	*0*	*1*	*2*	*3*
US	*5%*	*6%*	*7%*	*8%*
UK	*3%*	*4%*	*5%*	*4%*

(i) If the current exchange rate is £1 = \$1.5, find the fixed interest rate that would be paid on a plain-vanilla currency swap.

(ii) Determine the cash flows for the currency swap above if the principal is £100m, and show that the present value of the net flow is 0 for the firm receiving £.

Part VII

Application

Chapter 17

Portfolio Evaluation

17.1 Introduction

This must tie together some of the various components.

The basic issue will be to go through the investment process of selection, construction, investment and evalaution.

17.2 Portfolio Consturction

Could do this in a retrospective form

i.e. look at data in year 2000 to select a couple of different protfolios using the techniques descibed

one low risk, one high risk.

Can be related to two different people with different requirements

such as young and old.

17.3 Revision

Then a year later inspect these

Possibly revise

Then check again.

17.4 Longer Run

Bring up to the year 2005 to see how they perform.

17.5 Conclusion

Look at the issues that have been learnt.

Exercise 115 *Must do something similar as an exercise.*

Part VIII

Appendix

Chapter 18

Using Yahoo!

18.1 Introduction

In the recent past summary financial data could be obtained from daily newspapers or detailed financial data from subscritpion services. The situation now is that detailed data and research tools are widely available online. The sites that can be used include Yahoo, the BBC, etc.

This appendix will describe how to use the service at Yahoo and explore some of the information that it provides. The motivation for focussing on Yahoo is the extensive nature of the information and the fact that it is likely to be stable in format and content.

18.2 Basics

Yahoo can be accessed using either www.yahoo.com or www.yahoo.co.uk (or, equally, through a range of other national sites). From the Yahoo home page the Finance tab is selected. The difference between .com and .co.uk is the default information that is provided once the finance homepage is reached. The .com address will display US information whereas the .co.uk will provide UK information. The underlying information can be accessed from either site. The layout of the two pages is also different.

Both homepages display a summary of market information. This includes the current trading position of the leading stock exchange (the Dow for .com and London for .co.uk) and a summary of news stories that are relevant for finance. Tabs at the top of the page provide access to a variety of tools for investment include the management of a personal portfolio.

FOI

18.3 Symbols

The data in Yahooo finance are organised around the idea of a *symbol*. A symbol is an abrreviaed name for an asset. For example, stock in the Boeing Company taded on the New York Stock Exchange have the symbol BA. This uniquely identifies the stock. As a second example, stock in Barclays plc, a British bank, traded in London have the symbol BARC.L.

These symbols are obtained by entering the company name into the Get Quotes search. Be warned: the official company name may not match the company name in common use so it may take some time to find the correct company. One way to check is to read the Profile for the symbols that the Get Quotes search suggests. These can help identify the correct company.

When the name "Ford" is entered into the Get Quote search and the "Show all results for Ford" option is selected from the drop-down menu 165 alternatives are offered. Among the 45 stocks that are listed are theose given in Table 18.3. One of these – F – is the correct symbol for the Ford Motor Company. The others refer to special financial instruments issued by Ford while others represent stock traded on the worldwide exchanges (e.g. BRU is Brussels).

The Yahoo finance summary page for Ford Motor Company is reached by selecting F. This page provides a summary of traidng information, headlines, reports, and other materials. It also includes a one day chart of the stock price. A range of further options are listed in the left margin. For Ford all of these options are operational. This is not the case for the stock of smaller companies.

18.4 Research

The background information on the company

18.5 Historical Stock Prices

How to quickly find stock prices

18.6 Options

How to find option prices and interpretation

18.7 Creating a Portfolio

Describe how a portfolio is created and managed.