Part I

Single-Period Portfolio Choice and Asset Pricing

Chapter 1

Expected Utility and Risk Aversion

Asset prices are determined by investors' risk preferences and by the distributions of assets' risky future payments. Economists refer to these two bases of prices as investor "tastes" and the economy's "technologies" for generating asset returns. A satisfactory theory of asset valuation must consider how individuals allocate their wealth among assets having different future payments. This chapter explores the development of expected utility theory, the standard approach for modeling investor choices over risky assets. We first analyze the conditions that an individual's preferences must satisfy to be consistent with an expected utility function. We then consider the link between utility and risk aversion and how risk aversion leads to risk premia for particular assets. Our final topic examines how risk aversion affects an individual's choice between a risky and a risk-free asset.

Modeling investor choices with expected utility functions is widely used. However, significant empirical and experimental evidence has indicated that

individuals sometimes behave in ways inconsistent with standard forms of expected utility. These findings have motivated a search for improved models of investor preferences. Theoretical innovations both within and outside the expected utility paradigm are being developed, and examples of such advances are presented in later chapters of this book.

1.1 Preferences when Returns Are Uncertain

Economists typically analyze the price of a good or service by modeling the nature of its supply and demand. A similar approach can be taken to price an asset. As a starting point, let us consider the modeling of an investor's demand for an asset. In contrast to a good or service, an asset does not provide a current consumption benefit to an individual. Rather, an asset is a vehicle for saving. It is a component of an investor's financial wealth representing a claim on *future* consumption or purchasing power. The main distinction between assets is the difference in their future payoffs. With the exception of assets that pay a risk-free return, assets' payoffs are random. Thus, a theory of the demand for assets needs to specify investors' preferences over different, uncertain payoffs. In other words, we need to model how investors choose between assets that have different probability distributions of returns. In this chapter we assume an environment where an individual chooses among assets that have random payoffs at a single future date. Later chapters will generalize the situation to consider an individual's choices over multiple periods among assets paying returns at multiple future dates.

Let us begin by considering potentially relevant criteria that individuals might use to rank their preferences for different risky assets. One possible measure of the attractiveness of an asset is the average, or *expected value*, of its payoff. Suppose an asset offers a single random payoff at a particular

future date, and this payoff has a discrete distribution with n possible outcomes $(x_1, ..., x_n)$ and corresponding probabilities $(p_1, ..., p_n)$, where $\sum_{i=1}^{n} p_i = 1$ and $p_i \geq 0$.[1] Then the expected value of the payoff (or, more simply, the expected payoff) is $\bar{x} \equiv E[\widetilde{x}] = \sum_{i=1}^{n} p_i x_i$.

Is it logical to think that individuals value risky assets based solely on the assets' expected payoffs? This valuation concept was the prevailing wisdom until 1713, when Nicholas Bernoulli pointed out a major weakness. He showed that an asset's expected payoff was unlikely to be the only criterion that individuals use for valuation. He did it by posing the following problem which became known as the St. Petersberg paradox:

> Peter tosses a coin and continues to do so until it should land "heads" when it comes to the ground. He agrees to give Paul one ducat if he gets heads on the very first throw, two ducats if he gets it on the second, four if on the third, eight if on the fourth, and so on, so that on each additional throw the number of ducats he must pay is doubled.[2] Suppose we seek to determine Paul's expectation (of the payoff that he will receive).

Interpreting Paul's prize from this coin flipping game as the payoff of a risky asset, how much would he be willing to pay for this asset if he valued it based on its expected value? If the number of coin flips taken to first arrive at a heads is i, then $p_i = \left(\frac{1}{2}\right)^i$ and $x_i = 2^{i-1}$ so that the expected payoff equals

[1] As is the case in the following example, n, the number of possible outcomes, may be infinite.

[2] A ducat was a 3.5-gram gold coin used throughout Europe.

$$\begin{aligned} \bar{x} &= \sum_{i=1}^{\infty} p_i x_i = \tfrac{1}{2}1 + \tfrac{1}{4}2 + \tfrac{1}{8}4 + \tfrac{1}{16}8 + ... \\ &= \tfrac{1}{2}(1 + \tfrac{1}{2}2 + \tfrac{1}{4}4 + \tfrac{1}{8}8 + ... \\ &= \tfrac{1}{2}(1 + 1 + 1 + 1 + ... = \infty \end{aligned} \tag{1.1}$$

The "paradox" is that the expected value of this asset is infinite, but intuitively, most individuals would pay only a moderate, not infinite, amount to play this game. In a paper published in 1738, Daniel Bernoulli, a cousin of Nicholas's, provided an explanation for the St. Petersberg paradox by introducing the concept of *expected utility*.[3] His insight was that an individual's utility or "felicity" from receiving a payoff could differ from the size of the payoff and that people cared about the expected utility of an asset's payoffs, not the expected value of its payoffs. Instead of valuing an asset as $\overline{x} = \sum_{i=1}^{n} p_i x_i$, its value, V, would be

$$V \equiv E\left[U\left(\widetilde{x}\right)\right] = \textstyle\sum_{i=1}^{n} p_i U_i \tag{1.2}$$

where U_i is the utility associated with payoff x_i. Moreover, he hypothesized that the "utility resulting from any small increase in wealth will be inversely proportionate to the quantity of goods previously possessed." In other words, the greater an individual's wealth, the smaller is the added (or marginal) utility received from an additional increase in wealth. In the St. Petersberg paradox, prizes, x_i, go up at the same rate that the probabilities decline. To obtain a finite valuation, the trick is to allow the utility of prizes, U_i, to increase

[3] An English translation of Daniel Bernoulli's original Latin paper is printed in *Econometrica* (Bernoulli 1954). Another Swiss mathematician, Gabriel Cramer, offered a similar solution in 1728.

more slowly than the rate that probabilities decline. Hence, Daniel Bernoulli introduced the principle of a *diminishing marginal utility of wealth* (as expressed in his preceding quote) to resolve this paradox.

The first complete axiomatic development of expected utility is due to John von Neumann and Oskar Morgenstern (von Neumann and Morgenstern 1944). Von Neumann, a renowned physicist and mathematician, initiated the field of *game theory*, which analyzes strategic decision making. Morgenstern, an economist, recognized the field's economic applications and, together, they provided a rigorous basis for individual decision making under uncertainty. We now outline one aspect of their work, namely, to provide conditions that an individual's preferences must satisfy for these preferences to be consistent with an expected utility function.

Define a *lottery* as an asset that has a risky payoff and consider an individual's optimal choice of a lottery (risky asset) from a given set of different lotteries. All lotteries have possible payoffs that are contained in the set $\{x_1, ..., x_n\}$. In general, the elements of this set can be viewed as different, uncertain outcomes. For example, they could be interpreted as particular consumption levels (bundles of consumption goods) that the individual obtains in different states of nature or, more simply, different monetary payments received in different states of the world. A given lottery can be characterized as an ordered set of probabilities $P = \{p_1, ..., p_n\}$, where of course, $\sum_{i=1}^{n} p_i = 1$ and $p_i \geq 0$. A different lottery is characterized by another set of probabilities, for example, $P^* = \{p_1^*, ..., p_n^*\}$. Let $\succ$, $\prec$, and $\sim$ denote preference and indifference between lotteries.[4]

We will show that if an individual's preferences satisfy the following five conditions (axioms), then these preferences can be represented by a real-valued

[4]Specifically, if an individual prefers lottery P to lottery P^*, this can be denoted as $P \succ P^*$ or $P^* \prec P$. When the individual is indifferent between the two lotteries, this is written as $P \sim P^*$. If an individual prefers lottery P to lottery P^*or she is indifferent between lotteries P and P^*, this is written as $P \succeq P^*$ or $P^* \preceq P$.

utility function defined over a given lottery's probabilities, that is, an expected utility function $V(p_1, ..., p_n)$.

Axioms:

1) *Completeness*

For any two lotteries P^* and P, either $P^* \succ P$, or $P^* \prec P$, or $P^* \sim P$.

2) *Transitivity*

If $P^{**} \succeq P^*$and $P^* \succeq P$, then $P^{**} \succeq P$.

3) *Continuity*

If $P^{**} \succeq P^* \succeq P$, there exists some $\lambda \in [0, 1]$ such that $P^* \sim \lambda P^{**} + (1-\lambda)P$, where $\lambda P^{**} + (1-\lambda)P$ denotes a "compound lottery"; namely, with probability λ one receives the lottery P^{**} and with probability $(1-\lambda)$ one receives the lottery P.

These three axioms are analogous to those used to establish the existence of a real-valued utility function in standard consumer choice theory.[5] The fourth axiom is unique to expected utility theory and, as we later discuss, has important implications for the theory's predictions.

4) *Independence*

For any two lotteries P and P^*, $P^* \succ P$ if and only if for all $\lambda \in$ (0,1] and all P^{**}:

$$\lambda P^* + (1-\lambda)P^{**} \succ \lambda P + (1-\lambda)P^{**}$$

Moreover, for any two lotteries P and $P^\dagger$, $P \sim P^\dagger$ if and only if for all λ

[5] A primary area of microeconomics analyzes a consumer's optimal choice of multiple goods (and services) based on their prices and the consumer's budget contraint. In that context, utility is a function of the quantities of multiple goods consumed. References on this topic include (Kreps 1990), (Mas-Colell, Whinston, and Green 1995), and (Varian 1992). In contrast, the analysis of this chapter expresses utility as a function of the individual's wealth. In future chapters, we introduce multiperiod utility functions where utility becomes a function of the individual's overall consumption at multiple future dates. Financial economics typically bypasses the individual's problem of choosing among different consumption goods and focuses on how the individual chooses a total quantity of consumption at different points in time and different states of nature.

$\in(0,1]$ and all P^{**}:

$$\lambda P + (1-\lambda)P^{**} \sim \lambda P^{\dagger} + (1-\lambda)P^{**}$$

To better understand the meaning of the independence axiom, suppose that P^* is preferred to P. Now the choice between $\lambda P^* + (1-\lambda)P^{**}$ and $\lambda P + (1-\lambda)P^{**}$ is equivalent to a toss of a coin that has a probability $(1-\lambda)$ of landing "tails," in which case both compound lotteries are equivalent to P^{**}, and a probability λ of landing "heads," in which case the first compound lottery is equivalent to the single lottery P^* and the second compound lottery is equivalent to the single lottery P. Thus, the choice between $\lambda P^* + (1-\lambda)P^{**}$ and $\lambda P + (1-\lambda)P^{**}$ is equivalent to being asked, prior to the coin toss, if one would prefer P^* to P in the event the coin lands heads.

It would seem reasonable that should the coin land heads, we would go ahead with our original preference in choosing P^* over P. The independence axiom assumes that preferences over the two lotteries are independent of the way in which we obtain them.[6] For this reason, the independence axiom is also known as the "no regret" axiom. However, experimental evidence finds some systematic violations of this independence axiom, making it a questionable assumption for a theory of investor preferences. For example, the Allais paradox is a well-known choice of lotteries that, when offered to individuals, leads most to violate the independence axiom.[7] Machina (Machina 1987) summarizes violations of the independence axiom and reviews alternative approaches to modeling risk preferences. In spite of these deficiencies, the von Neumann-Morgenstern ex-

[6] In the context of standard consumer choice theory, λ would be interpreted as the amount (rather than probability) of a particular good or bundle of goods consumed (say C) and $(1-\lambda)$ as the amount of another good or bundle of goods consumed (say C^{**}). In this case, it would not be reasonable to assume that the choice of these different bundles is independent. This is due to some goods being substitutes or complements with other goods. Hence, the validity of the independence axiom is linked to outcomes being uncertain (risky), that is, the interpretation of λ as a probability rather than a deterministic amount.

[7] A similar example is given in Exercise 2 at the end of this chapter.

pected utility theory continues to be a useful and common approach to modeling investor preferences, though research exploring alternative paradigms is growing.[8]

The final axiom is similar to the independence and completeness axioms.

5) *Dominance*

Let P^1 be the compound lottery $\lambda_1 P^\ddagger + (1-\lambda_1)P^\dagger$ and P^2 be the compound lottery $\lambda_2 P^\ddagger + (1-\lambda_2)P^\dagger$. If $P^\ddagger \succ P^\dagger$, then $P^1 \succ P^2$ if and only if $\lambda_1 > \lambda_2$.

Given preferences characterized by the preceding axioms, we now show that the choice between any two (*or more*) arbitrary lotteries is that which has the higher (*highest*) expected utility.

The completeness axiom's ordering on lotteries naturally induces an ordering on the set of outcomes. To see this, define an "elementary" or "primitive" lottery, e_i, which returns outcome x_i with probability 1 and all other outcomes with probability zero, that is, $e_i = \{\mathrm{p}_1, ..., \mathrm{p}_{i-1}, \mathrm{p}_i, \mathrm{p}_{i+1}, ..., \mathrm{p}_n\} = \{0, ..., 0, 1, 0, ...0\}$ where $\mathrm{p}_i = 1$ and $\mathrm{p}_j = 0 \ \forall j \neq i$. Without loss of generality, suppose that the outcomes are ordered such that $e_n \succeq e_{n-1} \succeq ... \succeq e_1$. This follows from the completeness axiom for this case of n elementary lotteries. Note that this ordering of the elementary lotteries may not necessarily coincide with a ranking of the elements of x strictly by the size of their monetary payoffs, since the state of nature for which x_i is the outcome may differ from the state of nature for which x_j is the outcome, and these states of nature may have different effects on how an individual values the same monetary outcome. For example, x_i may be received in a state of nature when the economy is depressed, and monetary payoffs may be highly valued in this state of nature. In contrast, x_j may be received in a state of nature characterized by high economic expansion, and monetary payments may not be as highly valued. Therefore, it may be that

[8] This research includes "behavioral finance," a field that encompasses alternatives to both expected utility theory and market efficiency. An example of how a behavioral finance utility specification can impact asset prices will be presented in Chapter 15.

$e_i \succ e_j$ even if the monetary payment corresponding to x_i was less than that corresponding to x_j.

From the continuity axiom, we know that for each e_i, there exists a $U_i \in [0,1]$ such that

$$e_i \sim U_i e_n + (1 - U_i)e_1 \tag{1.3}$$

and for $i = 1$, this implies $U_1 = 0$ and for $i = n$, this implies $U_n = 1$. The values of the U_i weight the most and least preferred outcomes such that the individual is just indifferent between a combination of these polar payoffs and the payoff of x_i. The U_i can adjust for both differences in monetary payoffs and differences in the states of nature during which the outcomes are received.

Now consider a given arbitrary lottery, $P = \{p_1, ..., p_n\}$. This can be considered a compound lottery over the n elementary lotteries, where elementary lottery e_i is obtained with probability p_i. By the independence axiom, and using equation (1.3), the individual is indifferent between the compound lottery, P, and the following lottery, given on the right-hand side of the equation:

$$\begin{aligned} p_1e_1 + ... + p_ne_n \quad \sim \quad & p_1e_1 + ... + p_{i-1}e_{i-1} + p_i\left[U_ie_n + (1 - U_i)e_1\right] \\ & + p_{i+1}e_{i+1} + ... + p_ne_n \end{aligned} \tag{1.4}$$

where we have used the indifference relation in equation (1.3) to substitute for e_i on the right-hand side of (1.4). By repeating this substitution for all i, $i = 1, ..., n$, we see that the individual will be indifferent between P, given by the left-hand side of (1.4), and

$$p_1e_1 + ... + p_ne_n \sim \left(\sum_{i=1}^{n} p_iU_i\right) e_n + \left(1 - \sum_{i=1}^{n} p_iU_i\right) e_1 \tag{1.5}$$

Now define $\Lambda \equiv \sum_{i=1}^{n} p_i U_i$. Thus, we see that lottery P is equivalent to a compound lottery consisting of a Λ probability of obtaining elementary lottery e_n and a $(1-\Lambda)$ probability of obtaining elementary lottery e_1. In a similar manner, we can show that any other arbitrary lottery $P^* = \{p_1^*, ..., p_n^*\}$ is equivalent to a compound lottery consisting of a Λ^* probability of obtaining e_n and a $(1-\Lambda^*)$ probability of obtaining e_1, where $\Lambda^* \equiv \sum_{i=1}^{n} p_i^* U_i$.

Thus, we know from the dominance axiom that $P^* \succ P$ if and only if $\Lambda^* > \Lambda$, which implies $\sum_{i=1}^{n} p_i^* U_i > \sum_{i=1}^{n} p_i U_i$. So defining an expected utility function as

$$V(p_1, ..., p_n) = \sum_{i=1}^{n} p_i U_i \tag{1.6}$$

will imply that $P^* \succ P$ if and only if $V(p_1^*, ..., p_n^*) > V(p_1, ..., p_n)$.

The function given in equation (1.6) is known as von Neumann-Morgenstern expected utility. Note that it is linear in the probabilities and is unique up to a linear monotonic transformation.[9] This implies that the utility function has "cardinal" properties, meaning that it does not preserve preference orderings for all strictly increasing transformations.[10] For example, if $U_i = U(x_i)$, an individual's choice over lotteries will be the same under the transformation $aU(x_i) + b$, but not a nonlinear transformation that changes the "shape" of $U(x_i)$.

The von Neumann-Morgenstern expected utility framework may only partially explain the phenomenon illustrated by the St. Petersberg paradox. Suppose an individual's utility is given by the square root of a monetary payoff; that is, $U_i = U(x_i) = \sqrt{x_i}$. This is a monotonically increasing, concave function of

[9] The intuition for why expected utility is unique up to a linear transformation can be traced to equation (1.3). Here the derivation compares elementary lottery i in terms of the least and most preferred elementary lotteries. However, other bases for ranking a given lottery are possible.

[10] An "ordinal" utility function preserves preference orderings for *any* strictly increasing transformation, not just linear ones. The utility functions defined over multiple goods and used in standard consumer theory are ordinal measures.

x, which here is assumed to be simply a monetary amount (in units of ducats). Then the individual's expected utility of the St. Petersberg payoff is

$$\begin{aligned} V &= \sum_{i=1}^{n} p_i U_i = \sum_{i=1}^{\infty} \frac{1}{2^i} \sqrt{2^{i-1}} = \sum_{i=2}^{\infty} 2^{-\frac{i}{2}} \\ &= 2^{-\frac{2}{2}} + 2^{-\frac{3}{2}} + ... \\ &= \sum_{i=0}^{\infty} \left(\frac{1}{\sqrt{2}} \right)^i - 1 - \frac{1}{\sqrt{2}} = \frac{1}{2 - \sqrt{2}} \cong 1.707 \end{aligned} \tag{1.7}$$

which is finite. This individual would get the same expected utility from receiving a certain payment of $1.707^2 \cong 2.914$ ducats since $V = \sqrt{2.914}$ also gives expected (and actual) utility of 1.707. Hence, we can conclude that the St. Petersberg gamble would be worth 2.914 ducats to this square-root utility maximizer.

However, the reason that this is not a complete resolution of the paradox is that one can always construct a "super St. Petersberg paradox" where even expected utility is infinite. Note that in the regular St. Petersberg paradox, the probability of winning declines at rate 2^i, while the winning payoff increases at rate 2^i. In a super St. Petersberg paradox, we can make the winning payoff increase at a rate $x_i = U^{-1}(2^{i-1})$ and expected utility would no longer be finite. If we take the example of square-root utility, let the winning payoff be $x_i = 2^{2i-2}$; that is, $x_1 = 1$, $x_2 = 4$, $x_3 = 16$, and so on. In this case, the expected utility of the super St. Petersberg payoff by a square-root expected utility maximizer is

$$V = \sum_{i=1}^{n} p_i U_i = \sum_{i=1}^{\infty} \frac{1}{2^i} \sqrt{2^{2i-2}} = \infty \tag{1.8}$$

Should we be concerned that if we let the prizes grow quickly enough, we can get infinite expected utility (and valuations) for any chosen form of expected

utility function? Maybe not. One could argue that St. Petersberg games are unrealistic, particularly ones where the payoffs are assumed to grow rapidly. The reason is that any person offering this asset has finite wealth (even Bill Gates). This would set an upper bound on the amount of prizes that could feasibly be paid, making expected utility, and even the expected value of the payoff, finite.

The von Neumann-Morgenstern expected utility approach can be generalized to the case of a continuum of outcomes and lotteries having continuous probability distributions. For example, if outcomes are a possibly infinite number of purely monetary payoffs or consumption levels denoted by the variable x, a subset of the real numbers, then a generalized version of equation (1.6) is

$$V(F) = E[U(\widetilde{x})] = \int U(x)\, dF(x) \tag{1.9}$$

where $F(x)$ is a given lottery's cumulative distribution function over the payoffs, x.[11] Hence, the generalized lottery represented by the distribution function F is analogous to our previous lottery represented by the discrete probabilities $P = \{p_1, ..., p_n\}$.

Thus far, our discussion of expected utility theory has said little regarding an appropriate specification for the utility function, $U(x)$. We now turn to a discussion of how the form of this function affects individuals' risk preferences.

1.2 Risk Aversion and Risk Premia

As mentioned in the previous section, Daniel Bernoulli proposed that utility functions should display diminishing marginal utility; that is, $U(x)$ should be an increasing but concave function of wealth. He recognized that this concavity

[11] When the random payoff, $\widetilde{x}$, is absolutely continuous, then expected utility can be written in terms of the probability density function, $f(x)$, as $V(f) = \int U(x) f(x)\, dx$.

implies that an individual will be risk averse. By risk averse we mean that the individual would not accept a "fair" lottery (asset), where a fair or "pure risk" lottery is defined as one that has an expected value of zero. To see the relationship between fair lotteries and concave utility, consider the following example. Let there be a lottery that has a random payoff, $\widetilde{\varepsilon}$, where

$$\widetilde{\varepsilon} = \begin{cases} \varepsilon_1 \text{with probability } p \\ \varepsilon_2 \text{ with probability } 1-p \end{cases} \tag{1.10}$$

The requirement that it be a fair lottery restricts its expected value to equal zero:

$$E\left[\widetilde{\varepsilon}\right] = p\varepsilon_1 + (1-p)\varepsilon_2 = 0 \tag{1.11}$$

which implies $\varepsilon_1/\varepsilon_2 = -(1-p)/p$, or solving for p, $p = -\varepsilon_2/(\varepsilon_1 - \varepsilon_2)$. Of course, since $0 < p < 1$, ε_1 and ε_2 are of opposite signs.

Now suppose a von Neumann-Morgenstern expected utility maximizer whose current wealth equals W is offered the preceding lottery. Would this individual accept it; that is, would she place a positive value on this lottery?

If the lottery is accepted, expected utility is given by $E\left[U\left(W+\widetilde{\varepsilon}\right)\right]$. Instead, if it is not accepted, expected utility is given by $E\left[U\left(W\right)\right] = U\left(W\right)$. Thus, an individual's refusal to accept a fair lottery implies

$$U\left(W\right) > E\left[U\left(W+\widetilde{\varepsilon}\right)\right] = pU\left(W+\varepsilon_1\right) + (1-p)U\left(W+\varepsilon_2\right) \tag{1.12}$$

To show that this is equivalent to having a concave utility function, note that $U\left(W\right)$ can be rewritten as

$$U(W) = U\left(W + p\varepsilon_1 + (1-p)\varepsilon_2\right) \tag{1.13}$$

since $p\varepsilon_1 + (1-p)\varepsilon_2 = 0$ by the assumption that the lottery is fair. Rewriting inequality (1.12), we have

$$U\left(W + p\varepsilon_1 + (1-p)\varepsilon_2\right) > pU\left(W + \varepsilon_1\right) + (1-p)U\left(W + \varepsilon_2\right) \tag{1.14}$$

which is the definition of U being a concave function. A function is concave if a line joining any two points of the function lies entirely below the function. When $U(W)$ is concave, a line connecting the points $U(W+\varepsilon_2)$ to $U(W+\varepsilon_1)$ lies below $U(W)$ for all W such that $W+\varepsilon_2 < W < W+\varepsilon_1$. As shown in Figure 1.1, $pU(W+\varepsilon_1) + (1-p)U(W+\varepsilon_2)$ is exactly the point on this line directly below $U(W)$. This is clear by substituting $p = -\varepsilon_2/(\varepsilon_1 - \varepsilon_2)$. Note that when $U(W)$ is a continuous, second differentiable function, concavity implies that its second derivative, $U''(W)$, is less than zero.

To show the reverse, that concavity of utility implies the unwillingness to accept a fair lottery, we can use a result from statistics known as Jensen's inequality. If $U(\cdot)$ is some concave function and $\widetilde{x}$ is a random variable, then Jensen's inequality says that

$$E[U(\tilde{x})] < U(E[\tilde{x}]) \tag{1.15}$$

Therefore, substituting $\tilde{x} = W + \widetilde{\varepsilon}$ with $E[\widetilde{\varepsilon}] = 0$, we have

$$E\left[U(W + \widetilde{\varepsilon})\right] < U\left(E\left[W + \widetilde{\varepsilon}\right]\right) = U(W) \tag{1.16}$$

which is the desired result.

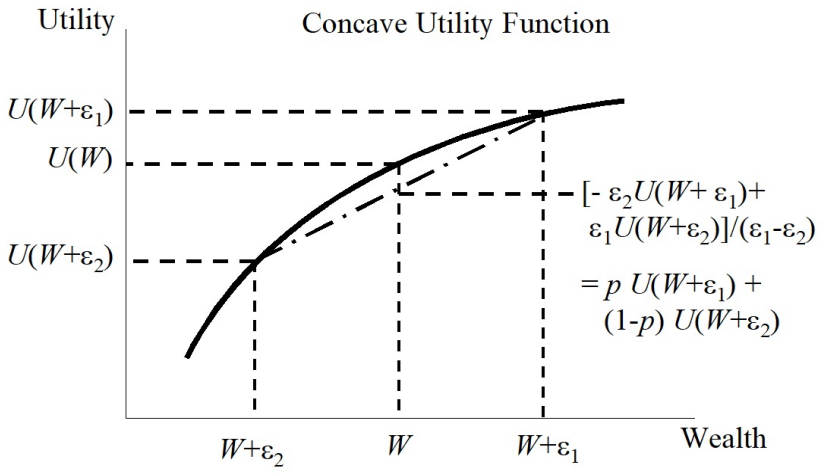


Figure 1.1: Fair Lotteries Lower Utility

We have defined risk aversion in terms of the individual's utility function.[12] Let us now consider how this aversion to risk can be quantified. This is done by defining a *risk premium*, the amount that an individual is willing to pay to avoid a risk.

Let π denote the individual's risk premium for a particular lottery, $\widetilde{\varepsilon}$. It can be likened to the maximum insurance payment an individual would pay to avoid a particular risk. John W. Pratt (Pratt 1964) defined the risk premium for lottery (asset) $\widetilde{\varepsilon}$ as

$$U(W-\pi) = E\left[U(W+\widetilde{\varepsilon})\right] \tag{1.17}$$

$W-\pi$ is defined as the *certainty equivalent* level of wealth associated with the

[12] Based on the same analysis, it is straightforward to show that if an individual strictly prefers a fair lottery, his utility function must be convex in wealth. Such an individual is said to be risk-loving. Similarly, an individual who is indifferent between accepting or refusing a fair lottery is said to be risk-neutral and must have utility that is a linear function of wealth.

lottery, $\widetilde{\varepsilon}$. Since utility is an increasing, concave function of wealth, Jensen's inequality ensures that π must be positive when $\widetilde{\varepsilon}$ is fair; that is, the individual would accept a level of wealth lower than her expected level of wealth following the lottery, $E[W+\widetilde{\varepsilon}]$, if the lottery could be avoided.

To analyze this Pratt (1964) risk premium, we continue to assume the individual is an expected utility maximizer and that $\widetilde{\varepsilon}$ is a fair lottery; that is, its expected value equals zero. Further, let us consider the case of $\widetilde{\varepsilon}$ being "small" so that we can study its effects by taking a Taylor series approximation of equation (1.17) around the point $\widetilde{\varepsilon}=0$ and $\pi=0$.[13] Expanding the left-hand side of (1.17) around $\pi=0$ gives

$$U(W-\pi) \cong U(W) - \pi U'(W) \tag{1.18}$$

and expanding the right-hand side of (1.17) around the zero mean of $\widetilde{\varepsilon}$ (and taking a three term expansion since $E[\widetilde{\varepsilon}]=0$ implies that a third term is necessary for a limiting approximation) gives

$$E[U(W+\widetilde{\varepsilon})] \cong E\left[U(W)+\widetilde{\varepsilon}U'(W)+\tfrac{1}{2}\widetilde{\varepsilon}^2U''(W)\right] \tag{1.19}$$

$$= U(W)+\tfrac{1}{2}\sigma^2U''(W)$$

where $\sigma^2 \equiv E\left[\widetilde{\varepsilon}^2\right]$ is the lottery's variance. Equating the results in (1.18) and (1.19), we have

$$\pi = -\tfrac{1}{2}\sigma^2\frac{U''(W)}{U'(W)} \equiv \tfrac{1}{2}\sigma^2R(W) \tag{1.20}$$

where $R(W) \equiv -U''(W)/U'(W)$ is the Pratt (1964)-Arrow (1971) measure of

[13] By describing the random variable $\widetilde{\varepsilon}$ as "small," we mean that its probability density is concentrated around its mean of 0.

absolute risk aversion. Note that the risk premium, π, depends on the uncertainty of the risky asset, σ^2, and on the individual's coefficient of absolute risk aversion. Since σ^2 and $U'(W)$ are both greater than zero, concavity of the utility function ensures that π must be positive.

From equation (1.20) we see that the concavity of the utility function, $U''(W)$, is insufficient to quantify the risk premium an individual is willing to pay, even though it is necessary and sufficient to indicate whether the individual is risk averse. In order to determine the risk premium, we also need the first derivative, $U'(W)$, which tells us the marginal utility of wealth. An individual may be very risk averse ($-U''(W)$ is large), but he may be unwilling to pay a large risk premium if he is poor since his marginal utility is high ($U'(W)$ is large).

To illustrate this point, consider the following *negative exponential* utility function:

$$U(W) = -e^{-bW}, b > 0 \tag{1.21}$$

Note that $U'(W) = be^{-bW} > 0$ and $U''(W) = -b^2e^{-bW} < 0$. Consider the behavior of a very wealthy individual, that is, one whose wealth approaches infinity:

$$\lim_{W\to\infty} U'(W) = \lim_{W\to\infty} U''(W) = 0 \tag{1.22}$$

As $W \to \infty$, the utility function is a flat line. Concavity disappears, which might imply that this very rich individual would be willing to pay very little for insurance against a random event, $\widetilde{\varepsilon}$, certainly less than a poor person with the same utility function. However, this is not true, because the marginal utility of wealth is also very small. This neutralizes the effect of smaller concavity.

Indeed,

$$R(W) = \frac{b^2 e^{-bW}}{be^{-bW}} = b \tag{1.23}$$

which is a constant. Thus, we can see why this utility function is sometimes referred to as a *constant absolute-risk-aversion* utility function.

If we want to assume that absolute risk aversion is declining in wealth, a necessary, though not sufficient, condition for this is that the utility function have a positive third derivative, since

$$\frac{\partial R(W)}{\partial W} = -\frac{U'''(W)U'(W) - [U''(W)]^2}{[U'(W)]^2} \tag{1.24}$$

Also, it can be shown that the coefficient of risk aversion contains all relevant information about the individual's risk preferences. To see this, note that

$$R(W) = -\frac{U''(W)}{U'(W)} = -\frac{\partial \left(\ln \left[U'(W)\right]\right)}{\partial W} \tag{1.25}$$

Integrating both sides of (1.25), we have

$$-\int R(W)dW = \ln[U'(W)] + c_1 \tag{1.26}$$

where c_1 is an arbitrary constant. Taking the exponential function of (1.26), one obtains

$$e^{-\int R(W)dW} = U'(W)e^{c_1} \tag{1.27}$$

Integrating once again gives

$$\int e^{-\int R(W)dW} dW = e^{c_1}U(W) + c_2 \tag{1.28}$$

where c_2 is another arbitrary constant. Because expected utility functions are unique up to a linear transformation, $e^{c_1}U(W) + c_1$ reflects the same risk preferences as $U(W)$. Hence, this shows one can recover the risk preferences of $U(W)$ from the function $R(W)$.

Relative risk aversion is another frequently used measure of risk aversion and is defined simply as

$$R_r(W) = WR(W) \tag{1.29}$$

In many applications in financial economics, an individual is assumed to have relative risk aversion that is constant for different levels of wealth. Note that this assumption implies that the individual's absolute risk aversion, $R(W)$, declines in direct proportion to increases in his wealth. While later chapters will discuss the widely varied empirical evidence on the size of individuals' relative risk aversions, one recent study based on individuals' answers to survey questions finds a median relative risk aversion of approximately 7.[14]

Let us now examine the coefficients of risk aversion for some utility functions that are frequently used in models of portfolio choice and asset pricing. *Power* utility can be written as

$$U(W) = \tfrac{1}{\gamma}W^{\gamma}, \gamma < 1 \tag{1.30}$$

[14] The mean estimate was lower, indicating a skewed distribution. Robert Barsky, Thomas Juster, Miles Kimball, and Matthew Shapiro (Barsky, Juster, Kimball, and Shapiro 1997) computed these estimates of relative risk aversion from a survey that asked a series of questions regarding whether the respondent would switch to a new job that had a 50-50 chance of doubling their lifetime income or decreasing their lifetime income by a proportion λ. By varying λ in the questions, they estimated the point where an individual would be indifferent between keeping their current job or switching. Essentially, they attempted to find λ^* such that $\frac{1}{2}U(2W) + \frac{1}{2}U(\lambda^* W) = U(W)$. Assuming utility displays constant relative risk aversion of the form $U(W) = W^{\gamma}/\gamma$, then the coefficient of relative risk aversion, $1-\gamma$, satisfies $2^{\gamma} + \lambda^{*\gamma} = 2$. The authors warn that their estimates of risk aversion may be biased upward if individuals attach nonpecuniary benefits to maintaining their current occupation. Interestingly, they confirmed that estimates of relative risk aversion tended to be lower for individuals who smoked, drank, were uninsured, held riskier jobs, and invested in riskier assets.

implying that $R(W) = -\frac{(\gamma-1)W^{\gamma-2}}{W^{\gamma-1}} = \frac{(1-\gamma)}{W}$ and, therefore, $R_r(W) = 1 - \gamma$. Hence, this form of utility is also known as *constant relative risk aversion*. *Logarithmic* utility is a limiting case of power utility. To see this, write the power utility function as $\frac{1}{\gamma}W^{\gamma} - \frac{1}{\gamma} = \frac{W^{\gamma}-1}{\gamma}$.[15] Next take the limit of this utility function as $\gamma \to 0$. Note that the numerator and denominator both go to zero, such that the limit is not obvious. However, we can rewrite the numerator in terms of an exponential and natural log function and apply L'Hôpital's rule to obtain

$$\lim_{\gamma\to 0}\frac{W^{\gamma}-1}{\gamma} = \lim_{\gamma\to 0}\frac{e^{\gamma\ln(W)}-1}{\gamma} = \lim_{\gamma\to 0}\frac{\ln(W)W^{\gamma}}{1} = \ln(W) \tag{1.31}$$

Thus, logarithmic utility is equivalent to power utility with $\gamma = 0$, or a coefficient of relative risk aversion of unity: $R(W) = -\frac{W^{-2}}{W^{-1}} = \frac{1}{W}$ and $R_r(W) = 1$.

Quadratic utility takes the form

$$U(W) = W - \tfrac{b}{2}W^2, b > 0 \tag{1.32}$$

Note that the marginal utility of wealth is $U'(W) = 1 - bW$ and is positive only when $b < \frac{1}{W}$. Thus, this utility function makes sense (in that more wealth is preferred to less) only when $W < \frac{1}{b}$. The point of maximum utility, $\frac{1}{b}$, is known as the "bliss point." We have $R(W) = \frac{b}{1-bW}$ and $R_r(W) = \frac{bW}{1-bW}$.

Hyperbolic absolute-risk-aversion (HARA) utility is a generalization of all of the aforementioned utility functions. It can be written as

$$U(W) = \frac{1-\gamma}{\gamma}\left(\frac{\alpha W}{1-\gamma} + \beta\right)^{\gamma} \tag{1.33}$$

[15] Recall that we can do this because utility functions are unique up to a linear transformation.

subject to the restrictions $\gamma \neq 1$, $\alpha > 0$, $\frac{\alpha W}{1-\gamma} + \beta > 0$, and $\beta = 1$ if $\gamma = -\infty$. Thus, $R(W) = \left(\frac{W}{1-\gamma} + \frac{\beta}{\alpha}\right)^{-1}$. Since $R(W)$ must be > 0, it implies $\beta > 0$ when $\gamma > 1$. $R_r(W) = W\left(\frac{W}{1-\gamma} + \frac{\beta}{\alpha}\right)^{-1}$. HARA utility nests constant absolute risk aversion ($\gamma = -\infty$, $\beta = 1$), constant relative risk aversion ($\gamma < 1$, $\beta = 0$), and quadratic ($\gamma = 2$) utility functions. Thus, depending on the parameters, it is able to display constant absolute risk aversion or relative risk aversion that is increasing, decreasing, or constant. We will revisit HARA utility in future chapters as it can be an analytically convenient assumption for utility when deriving an individual's intertemporal consumption and portfolio choices.

Pratt's definition of a risk premium in equation (1.17) is commonly used in the insurance literature because it can be interpreted as the payment that an individual is willing to make to insure against a particular risk. However, in the field of financial economics, a somewhat different definition is often employed. Financial economists seek to understand how the risk of an asset's payoff determines the asset's rate of return. In this context, an asset's risk premium is defined as its expected rate of return in excess of the risk-free rate of return. This alternative concept of a risk premium was used by Kenneth Arrow (Arrow 1971), who independently derived a coefficient of risk aversion that is identical to Pratt's measure. Let us now outline Arrow's approach. Suppose that an asset (lottery), $\widetilde{\varepsilon}$, has the following payoffs and probabilities (which could be generalized to other types of fair payoffs):

$$\widetilde{\varepsilon} = \begin{cases} +\epsilon \text{ with probability } \frac{1}{2} \\ -\epsilon \text{ with probability } \frac{1}{2} \end{cases} \tag{1.34}$$

where $\epsilon \geq 0$. Note that, as before, $E[\widetilde{\varepsilon}] = 0$. Now consider the following question. By how much should we change the expected value (return) of the asset, by changing the probability of winning, in order to make the individual

indifferent between taking and not taking the risk? If p is the probability of winning, we can define the risk premium as

$$\theta = prob\left(\widetilde{\varepsilon} = +\epsilon\right) - prob\left(\widetilde{\varepsilon} = -\epsilon\right) = p - (1-p) = 2p - 1 \tag{1.35}$$

Therefore, from (1.35) we have

$$\begin{array}{c} prob\left(\widetilde{\varepsilon} = +\epsilon\right) \equiv p = \frac{1}{2}(1+\theta) \\ prob\left(\widetilde{\varepsilon} = -\epsilon\right) \equiv 1 - p = \frac{1}{2}(1-\theta) \end{array} \tag{1.36}$$

These new probabilities of winning and losing are equal to the old probabilities, $\frac{1}{2}$, plus half of the increment, θ. Thus, the premium, θ, that makes the individual indifferent between accepting and refusing the asset is

$$U(W) = \frac{1}{2}(1+\theta)U(W+\epsilon) + \frac{1}{2}(1-\theta)U(W-\epsilon) \tag{1.37}$$

Taking a Taylor series approximation around $\epsilon = 0$ gives

$$\begin{array}{rcl} U(W) & = & \frac{1}{2}(1+\theta)\left[U(W) + \epsilon U'(W) + \frac{1}{2}\epsilon^2 U''(W)\right] \\ & & + \frac{1}{2}(1-\theta)\left[U(W) - \epsilon U'(W) + \frac{1}{2}\epsilon^2 U''(W)\right] \\ & = & U(W) + \epsilon\theta U'(W) + \frac{1}{2}\epsilon^2 U''(W) \end{array} \tag{1.38}$$

Rearranging (1.38) implies

$$\theta = \tfrac{1}{2}\epsilon R(W) \tag{1.39}$$

which, as before, is a function of the coefficient of absolute risk aversion. Note that the Arrow premium, θ, is in terms of a probability, while the Pratt measure, π, is in units of a monetary payment. If we multiply θ by the monetary payment

received, ϵ, then equation (1.39) becomes

$$\epsilon\theta = \tfrac{1}{2}\epsilon^2 R(W) \tag{1.40}$$

Since ϵ^2 is the variance of the random payoff, $\widetilde{\varepsilon}$, equation (1.40) shows that the Pratt and Arrow measures of risk premia are equivalent. Both were obtained as a linearization of the true function around $\widetilde{\varepsilon} = 0$.

The results of this section showed how risk aversion depends on the shape of an individual's utility function. Moreover, it demonstrated that a risk premium, equal to either the payment an individual would make to avoid a risk or the individual's required excess rate of return on a risky asset, is proportional to the individual's Pratt-Arrow coefficient of absolute risk aversion.

1.3 Risk Aversion and Portfolio Choice

Having developed the concepts of risk aversion and risk premiums, we now consider the relation between risk aversion and an individual's portfolio choice in a single period context. While the portfolio choice problem that we analyze is very simple, many of its insights extend to the more complex environments that will be covered in later chapters of this book. We shall demonstrate that absolute and relative risk aversion play important roles in determining how portfolio choices vary with an individual's level of wealth. Moreover, we show that when given a choice between a risk-free asset and a risky asset, a risk-averse individual always chooses at least some positive investment in the risky asset if it pays a positive risk premium.

The model's assumptions are as follows. Assume there is a riskless security that pays a rate of return equal to r_f. In addition, for simplicity, suppose there is just one risky security that pays a stochastic rate of return equal to $\widetilde{r}$. Also,

let W_0 be the individual's initial wealth, and let A be the dollar amount that the individual invests in the risky asset at the beginning of the period. Thus, $W_0 - A$ is the initial investment in the riskless security. Denoting the individual's end-of-period wealth as $\tilde{W}$, it satisfies

$$\begin{aligned} \tilde{W} &= (W_0 - A)(1 + r_f) + A(1 + \tilde{r}) \\ &= W_0(1 + r_f) + A(\tilde{r} - r_f) \end{aligned} \tag{1.41}$$

Note that in the second line of equation (1.41), the first term is the individual's return on wealth when the entire portfolio is invested in the risk-free asset, while the second term is the difference in return gained by investing A dollars in the risky asset.

We assume that the individual cares only about consumption at the end of this single period. Therefore, maximizing end-of-period consumption is equivalent to maximizing end-of-period wealth. Assuming that the individual is a von Neumann-Morgenstern expected utility maximizer, she chooses her portfolio by maximizing the expected utility of end-of-period wealth:

$$\max_A E[U(\tilde{W})] = \max_A E\left[U\left(W_0(1 + r_f) + A(\tilde{r} - r_f)\right)\right] \tag{1.42}$$

The solution to the individual's problem in (1.42) must satisfy the following first-order condition with respect to A:

$$E\left[U'\left(\tilde{W}\right)(\tilde{r} - r_f)\right] = 0 \tag{1.43}$$

This condition determines the amount, A, that the individual invests in the

risky asset.[16] Consider the special case in which the expected rate of return on the risky asset equals the risk-free rate. In that case, $A = 0$ satisfies the first-order condition. To see this, note that when $A = 0$, then $\tilde{W} = W_0(1 + r_f)$ and, therefore, $U'\left(\tilde{W}\right) = U'(W_0(1 + r_f))$ are nonstochastic. Hence, $E\left[U'\left(\tilde{W}\right)(\tilde{r} - r_f)\right] = U'(W_0(1 + r_f))E[\tilde{r} - r_f] = 0$. This result is reminiscent of our earlier finding that a risk-averse individual would not choose to accept a fair lottery. Here, the fair lottery is interpreted as a risky asset that has an expected rate of return just equal to the risk-free rate.

Next, consider the case in which $E[\tilde{r}] - r_f > 0$. Clearly, $A = 0$ would not satisfy the first-order condition, because $E\left[U'\left(\tilde{W}\right)(\tilde{r} - r_f)\right] = U'(W_0(1 + r_f))E[\tilde{r} - r_f] > 0$ when $A = 0$. Rather, when $E[\tilde{r}] - r_f > 0$, condition (1.43) is satisfied only when $A > 0$. To see this, let r^h denote a realization of $\tilde{r}$ such that it exceeds r_f, and let W^h be the corresponding level of $\tilde{W}$. Also, let r^l denote a realization of $\tilde{r}$ such that it is lower than r_f, and let W^l be the corresponding level of $\tilde{W}$. Obviously, $U'(W^h)(r^h - r_f) > 0$ and $U'(W^l)(r^l - r_f) < 0$. For $U'\left(\tilde{W}\right)(\tilde{r} - r_f)$ to average to zero for all realizations of $\tilde{r}$, it must be the case that $W^h > W^l$ so that $U'\left(W^h\right) < U'\left(W^l\right)$ due to the concavity of the utility function. This is because $E[\tilde{r}] - r_f > 0$, so the average realization of r^h is farther above r_f than the average realization of r^l is below r_f. Therefore, to make $U'\left(\tilde{W}\right)(\tilde{r} - r_f)$ average to zero, the positive $(r^h - r_f)$ terms need to be given weights, $U'\left(W^h\right)$, that are smaller than the weights, $U'(W^l)$, that multiply the negative $(r^l - r_f)$ realizations. This can occur only if $A > 0$ so that $W^h > W^l$. The implication is that an individual will always hold at least some positive amount of the risky asset if its expected rate of return exceeds the risk-free rate.[17]

[16] The second order condition for a maximum, $E\left[U''\left(\tilde{W}\right)(\tilde{r} - r_f)^2\right] \leq 0$, is satisfied because $U''\left(\tilde{W}\right) \leq 0$ due to the assumed concavity of the utility function.

[17] Related to this is the notion that a risk-averse expected utility maximizer should accept a small lottery with a positive expected return. In other words, such an individual should be close to risk-neutral for small-scale bets. However, Matthew Rabin and Richard Thaler (Rabin and Thaler 2001) claim that individuals frequently reject lotteries (gambles) that are

Now, we can go further and explore the relationship between A and the individual's initial wealth, W_0. Using the envelope theorem, we can differentiate the first-order condition to obtain[18]

$$E\left[U''(\tilde{W})(\tilde{r}-r_f)(1+r_f)\right]dW_0 + E\left[U''(\tilde{W})(\tilde{r}-r_f)^2\right]dA = 0 \tag{1.44}$$

or

$$\frac{dA}{dW_0} = \frac{(1+r_f)E\left[U''(\tilde{W})(\tilde{r}-r_f)\right]}{-E\left[U''(\tilde{W})(\tilde{r}-r_f)^2\right]} \tag{1.45}$$

The denominator of (1.45) is positive because concavity of the utility function ensures that $U''(\tilde{W})$ is negative. Therefore, the sign of the expression depends on the numerator, which can be of either sign because realizations of $(\tilde{r}-r_f)$ can turn out to be both positive and negative.

To characterize situations in which the sign of (1.45) can be determined, let

modest in size yet have positive expected returns. From this they argue that concave expected utility is not a plausible model for predicting an individual's choice of small-scale risks.

[18] The envelope theorem is used to analyze how the maximized value of the objective function and the control variable change when one of the model's parameters changes. In our context, define $f(A,W_0) \equiv E\left[U\left(\widetilde{W}\right)\right]$ and let the function $v(W_0) = \max_A f(A,W_0)$ be the maximized value of the objective function when the control variable, A, is optimally chosen. Also define $A(W_0)$ as the value of A that maximizes f for a given value of the initial wealth parameter W_0. Now let us first consider how the maximized value of the objective function changes when we change the parameter W_0. We do this by differentiating $v(W_0)$ with respect to W_0 by applying the chain rule to obtain $\frac{dv(W_0)}{dW_0} = \frac{\partial f(A,W_0)}{\partial A}\frac{dA(W_0)}{dW_0} + \frac{\partial f(A(W_0),W_0)}{\partial W_0}$. However, note that $\frac{\partial f(A,W_0)}{\partial A} = 0$ since this is the first-order condition for a maximum, and it must hold when at the maximum. Hence, this derivative simplifies to $\frac{dv(W_0)}{dW_0} = \frac{\partial f(A(W_0),W_0)}{\partial W_0}$. Thus, the first envelope theorem result is that the derivative of the maximized value of the objective function with respect to a parameter is just the partial derivative with respect to that parameter. Second, consider how the optimal value of the control variable, $A(W_0)$, changes when the parameter W_0 changes. We can derive this relationship by differentiating the first-order condition $\partial f(A(W_0),W_0)/\partial A = 0$ with respect to W_0. Again applying the chain rule to the first-order condition, one obtains $\frac{\partial(\partial f(A(W_0),W_0)/\partial A)}{\partial W_0} = 0 = \frac{\partial^2 f(A(W_0),W_0)}{\partial A^2}\frac{dA(W_0)}{dW_0} + \frac{\partial^2 f(A(W_0),W_0)}{\partial A \partial W_0}$. Rearranging gives us $\frac{dA(W_0)}{dW_0} = -\frac{\partial^2 f(A(W_0),W_0)}{\partial A \partial W_0} \Big/ \frac{\partial^2 f(A(W_0),W_0)}{\partial A^2}$, which is equation (1.45).

us first consider the case where the individual has absolute risk aversion that is decreasing in wealth. As before, let r^h denote a realization of $\tilde{r}$ such that it exceeds r_f, and let W^h be the corresponding level of $\tilde{W}$. Then for $A \geqslant 0$, we have $W^h \geqslant W_0(1+r_f)$. If absolute risk aversion is decreasing in wealth, this implies

$$R\left(W^h\right) \leqslant R\left(W_0(1+r_f)\right) \tag{1.46}$$

where, as before, $R(W) = -U''(W)/U'(W)$. Multiplying both terms of (1.46) by $-U'(W^h)(r^h - r_f)$, which is a negative quantity, the inequality sign changes:

$$U''(W^h)(r^h - r_f) \geqslant -U'(W^h)(r^h - r_f)R\left(W_0(1+r_f)\right) \tag{1.47}$$

Next, we again let r^l denote a realization of $\tilde{r}$ that is lower than r_f and define W^l to be the corresponding level of $\tilde{W}$. Then for $A \geqslant 0$, we have $W^l \leqslant W_0(1+r_f)$. If absolute risk aversion is decreasing in wealth, this implies

$$R(W^l) \geqslant R\left(W_0(1+r_f)\right) \tag{1.48}$$

Multiplying (1.48) by $-U'(W^l)(r^l - r_f)$, which is positive, so that the sign of (1.48) remains the same, we obtain

$$U''(W^l)(r^l - r_f) \geqslant -U'(W^l)(r^l - r_f)R\left(W_0(1+r_f)\right) \tag{1.49}$$

Notice that inequalities (1.47) and (1.49) are of the same form. The inequality holds whether the realization is $\tilde{r} = r^h$ or $\tilde{r} = r^l$. Therefore, if we take expectations over all realizations, where $\tilde{r}$ can be either higher than or lower than r_f, we obtain

$$E\left[U''(\tilde{W})(\tilde{r}-r_f)\right] \gtrless -E\left[U'(\tilde{W})(\tilde{r}-r_f)\right] R\left(W_0(1+r_f)\right) \tag{1.50}$$

Since the first term on the right-hand side is just the first-order condition, inequality (1.50) reduces to

$$E\left[U''(\tilde{W})(\tilde{r}-r_f)\right] \gtrless 0 \tag{1.51}$$

Thus, the first conclusion that can be drawn is that declining absolute risk aversion implies $dA/dW_0 > 0$; that is, the individual invests an increasing amount of wealth in the risky asset for larger amounts of initial wealth. For two individuals with the same utility function but different initial wealths, the wealthier one invests a greater dollar amount in the risky asset if utility is characterized by decreasing absolute risk aversion. While not shown here, the opposite is true, namely, that the wealthier individual invests a smaller dollar amount in the risky asset if utility is characterized by increasing absolute risk aversion.

Thus far, we have not said anything about the *proportion* of initial wealth invested in the risky asset. To analyze this issue, we need the concept of relative risk aversion. Define

$$\eta \equiv \frac{dA}{dW_0}\frac{W_0}{A} \tag{1.52}$$

which is the elasticity measuring the proportional increase in the risky asset for an increase in initial wealth. Adding $1-\frac{A}{A}$ to the right-hand side of (1.52) gives

$$\eta = 1 + \frac{(dA/dW_0)W_0 - A}{A} \tag{1.53}$$

Substituting the expression dA/dW_0 from equation (1.45), we have

$$\eta = 1 + \frac{W_0(1+r_f)E\left[U''(\tilde{W})(\tilde{r}-r_f)\right] + AE\left[U''(\tilde{W})(\tilde{r}-r_f)^2\right]}{-AE\left[U''(\tilde{W})(\tilde{r}-r_f)^2\right]} \tag{1.54}$$

Collecting terms in $U''(\tilde{W})(\tilde{r}-r_f)$, this can be rewritten as

$$\begin{aligned} \eta &= 1 + \frac{E\left[U''(\tilde{W})(\tilde{r}-r_f)\{W_0(1+r_f)+A(\tilde{r}-r_f)\}\right]}{-AE\left[U''(\tilde{W})(\tilde{r}-r_f)^2\right]} \\ &= 1 + \frac{E\left[U''(\tilde{W})(\tilde{r}-r_f)\tilde{W}\right]}{-AE\left[U''(\tilde{W})(\tilde{r}-r_f)^2\right]} \end{aligned} \tag{1.55}$$

The denominator is always positive. Therefore, we see that the elasticity, η, is greater than one, so that the individual invests proportionally more in the risky asset with an increase in wealth, if $E\left[U''(\tilde{W})(\tilde{r}-r_f)\tilde{W}\right] \geqslant 0$. Can we relate this to the individual's risk aversion? The answer is yes and the derivation is almost exactly the same as that just given.

Consider the case where the individual has *relative* risk aversion that is decreasing in wealth. Let r^h denote a realization of $\tilde{r}$ such that it exceeds r_f, and let W^h be the corresponding level of $\tilde{W}$. Then for $A \geqslant 0$, we have $W^h \geqslant W_0(1+r_f)$. If relative risk aversion, $R_r(W) \equiv WR(W)$, is decreasing in wealth, this implies

$$W^h R(W^h) \leqslant W_0(1+r_f)R\left(W_0(1+r_f)\right) \tag{1.56}$$

Multiplying both terms of (1.56) by $-U'(W^h)(r^h - r_f)$, which is a negative quantity, the inequality sign changes:

$$W^h U''(W^h)(r^h - r_f) \geqslant -U'(W^h)(r^h - r_f)W_0(1 + r_f)R\left(W_0(1 + r_f)\right) \quad (1.57)$$

Next, let r^l denote a realization of $\tilde{r}$ such that it is lower than r_f, and let W^l be the corresponding level of $\tilde{W}$. Then for $A \geqslant 0$, we have $W^l \leqslant W_0(1 + r_f)$. If relative risk aversion is decreasing in wealth, this implies

$$W^l R(W^l) \geqslant W_0(1 + r_f)R\left(W_0(1 + r_f)\right) \quad (1.58)$$

Multiplying (1.58) by $-U'(W^l)(r^l - r_f)$, which is positive, so that the sign of (1.58) remains the same, we obtain

$$W^l U''(W^l)(r^l - r_f) \geqslant -U'(W^l)(r^l - r_f)W_0(1 + r_f)R\left(W_0(1 + r_f)\right) \quad (1.59)$$

Notice that inequalities (1.57) and (1.59) are of the same form. The inequality holds whether the realization is $\tilde{r} = r^h$ or $\tilde{r} = r^l$. Therefore, if we take expectations over all realizations, where $\tilde{r}$ can be either higher than or lower than r_f, we obtain

$$E\left[\tilde{W}U''(\tilde{W})(\tilde{r} - r_f)\right] \geqslant -E\left[U'(\tilde{W})(\tilde{r} - r_f)\right] W_0(1+r_f)R(W_0(1+r_f)) \quad (1.60)$$

Since the first term on the right-hand side is just the first-order condition, inequality (1.60) reduces to

$$E\left[\tilde{W}U''(\tilde{W})(\tilde{r} - r_f)\right] \geqslant 0 \quad (1.61)$$

Thus, we see that an individual with decreasing relative risk aversion has $\eta > 1$ and invests proportionally more in the risky asset as wealth increases. The opposite is true for increasing relative risk aversion: $\eta < 1$ so that this individual invests proportionally less in the risky asset as wealth increases. The following table provides another way of writing this section's main results.

Risk Aversion	**Investment Behavior**
Decreasing Absolute	$\frac{\partial A}{\partial W_0} > 0$
Constant Absolute	$\frac{\partial A}{\partial W_0} = 0$
Increasing Absolute	$\frac{\partial A}{\partial W_0} < 0$
Decreasing Relative	$\frac{\partial A}{\partial W_0} > \frac{A}{W_0}$
Constant Relative	$\frac{\partial A}{\partial W_0} = \frac{A}{W_0}$
Increasing Relative	$\frac{\partial A}{\partial W_0} < \frac{A}{W_0}$

A point worth emphasizing is that absolute risk aversion indicates how the investor's dollar amount in the risky asset changes with changes in initial wealth, whereas relative risk aversion indicates how the investor's portfolio proportion (or portfolio weight) in the risky asset, A/W_0, changes with changes in initial wealth.

1.4 Summary

This chapter is a first step toward understanding how an individual's preferences toward risk affect his portfolio behavior. It was shown that if an individual's risk preferences satisfied specific plausible conditions, then her behavior could be represented by a von Neumann-Morgenstern expected utility function. In turn, the shape of the individual's utility function determines a measure of risk aversion that is linked to two concepts of a risk premium. The first one is the monetary payment that the individual is willing to pay to avoid a risk, an example being a premium paid to insure against a property/casualty loss. The

second is the rate of return in excess of a riskless rate that the individual requires to hold a risky asset, which is the common definition of a security risk premium used in the finance literature. Finally, it was shown how an individual's absolute and relative risk aversion affect his choice between a risky and risk-free asset. In particular, individuals with decreasing (*increasing*) relative risk aversion invest proportionally more (*less*) in the risky asset as their wealth increases. Though based on a simple single-period, two-asset portfolio choice model, this insight generalizes to the more complex portfolio choice problems that will be studied in later chapters.

1.5 Exercises

1. Suppose there are two lotteries $P = \{p_1, ..., p_n\}$ and $P^* = \{p_1^*, ..., p_n^*\}$. Let $V(p_1, ..., p_n) = \sum_{i=1}^{n} p_i U_i$ be an individual's expected utility function defined over these lotteries. Let $W(p_1, ..., p_n) = \sum_{i=1}^{n} p_i Q_i$ where $Q_i = a + bU_i$ and a and b are constants. If $P^* \succ P$, so that $V(p_1^*, ..., p_n^*) > V(p_1, ..., p_n)$, must it be the case that $W(p_1^*, ..., p_n^*) > W(p_1, ..., p_n)$? In other words, is W also a valid expected utility function for the individual? Are there any restrictions needed on a and b for this to be the case?

2. (Allais paradox) Asset A pays \$1,500 with certainty, while asset B pays \$2,000 with probability 0.8 or \$100 with probability 0.2. If offered the choice between asset A or B, a particular individual would choose asset A. Suppose, instead, that the individual is offered the choice between asset C and asset D. Asset C pays \$1,500 with probability 0.25 or \$100 with probability 0.75, while asset D pays \$2,000 with probability 0.2 or \$100 with probability 0.8. If asset D is chosen, show that the individual's preferences violate the independence axiom.

3. Verify that the HARA utility function in equation (1.33) becomes the constant absolute-risk-aversion utility function when $\beta = 1$ and $\gamma = -\infty$. Hint: recall that $e^a = \lim_{x \longrightarrow \infty} \left(1 + \frac{a}{x}\right)^x$.

4. Consider the individual's portfolio choice problem given in equation (1.42). Assume $U(W) = \ln(W)$ and the rate of return on the risky asset equals $\widetilde{r} = \begin{cases} 4r_f & \text{with probability } \frac{1}{2} \\ -r_f & \text{with probability } \frac{1}{2} \end{cases}$. Solve for the individual's proportion of initial wealth invested in the risky asset, A/W_0.

5. An expected-utility-maximizing individual has constant relative-risk-aversion utility, $U(W) = W^\gamma/\gamma$, with relative risk-aversion coefficient of $\gamma = -1$. The individual currently owns a product that has a probability p of failing, an event that would result in a loss of wealth that has a present value equal to L. With probability $1 - p$, the product will not fail and no loss will result. The individual is considering whether to purchase an extended warranty on this product. The warranty costs C and would insure the individual against loss if the product fails. Assuming that the cost of the warranty exceeds the expected loss from the product's failure, determine the individual's level of wealth at which she would be just indifferent between purchasing or not purchasing the warranty.

6. In the context of the portfolio choice problem in equation (1.42), show that an individual with increasing relative risk aversion invests proportionally less in the risky asset as her initial wealth increases.

7. Consider the following four assets whose payoffs are as follows:

$$\text{Asset A} = \begin{cases} X \text{ with probability } p_x \\ 0 \text{ with probability } 1 - p_x \end{cases} \quad \text{Asset B} = \begin{cases} Y \text{ with probability } p_y \\ 0 \text{ with probability } 1 - p_y \end{cases}$$

$$\text{Asset C} = \begin{cases} X \text{ with probability } \alpha p_x \\ 0 \text{ with probability } 1 - \alpha p_x \end{cases} \qquad \text{Asset D} = \begin{cases} Y \text{ with probability } \alpha p_y \\ 0 \text{ with probability } 1 - \alpha p_y \end{cases}$$

where $0 < X < Y$, $p_y < p_x$, $p_x X < p_y Y$, and $\alpha \in (0, 1)$.

a. When given the choice of asset C versus asset D, an individual chooses asset C. Could this individual's preferences be consistent with von Neumann-Morgenstern expected utility theory? Explain why or why not.

b. When given the choice of asset A versus asset B, an individual chooses asset A. This same individual, when given the choice between asset C and asset D, chooses asset D. Could this individual's preferences be consistent with von Neumann-Morgenstern expected utility theory? Explain why or why not.

8. An individual has expected utility of the form

$$E\left[U\left(\widetilde{W}\right)\right] = E\left[-e^{-b\widetilde{W}}\right]$$

where $b > 0$. The individual's wealth is normally distributed as $N\left(\overline{W}, \sigma_W^2\right)$. What is this individual's *certainty equivalent* level of wealth?

Chapter 2

Mean-Variance Analysis

The preceding chapter studied an investor's choice between a risk-free asset and a single risky asset. This chapter adds realism by giving the investor the opportunity to choose among multiple risky assets. As a University of Chicago graduate student, Harry Markowitz wrote a path-breaking article on this topic (Markowitz 1952).[1] Markowitz's insight was to recognize that, in allocating wealth among various risky assets, a risk-averse investor should focus on the expectation and the risk of her combined portfolio's return, a return that is affected by the individual assets' diversification possibilities. Because of diversification, the attractiveness of a particular asset when held in a portfolio can differ from its appeal when it is the sole asset held by an investor.

Markowitz proxied the risk of a portfolio's return by the variance of its return. Of course, the variance of an investor's total portfolio return depends on the return variances of the individual assets included in the portfolio. But portfolio return variance also depends on the covariances of the individual as-

[1] His work on portfolio theory, of which this article was the beginning, won him a share of the Nobel prize in economics in 1990. Initially, the importance of his work was not widely recognized. Milton Friedman, a member of Markowitz's doctoral dissertation committee and later also a Nobel laureate, questioned whether the work met the requirements for an economics Ph.D. See (Bernstein 1992).

sets' returns. Hence, in selecting an optimal portfolio, the investor needs to consider how the comovement of individual assets' returns affects diversification possibilities.

A rational investor would want to choose a portfolio of assets that efficiently trades off higher expected return for lower variance of return. Interestingly, not all portfolios that an investor can create are efficient in this sense. Given the expected returns and covariances of returns on individual assets, Markowitz solved the investor's problem of constructing an efficient portfolio. His work has had an enormous impact on the theory and practice of portfolio management and asset pricing.

Intuitively, it makes sense that investors would want their wealth to earn a high average return with as little variance as possible. However, in general, an individual who maximizes expected utility may care about moments of the distribution of wealth in addition to its mean and variance.[2] Though Markowitz's mean-variance analysis fails to consider the effects of these other moments, in later chapters of this book we will see that his model's insights can be generalized to more complicated settings.

The next section outlines the assumptions on investor preferences and the distribution of asset returns that would allow us to simplify the investor's portfolio choice problem to one that considers only the mean and variance of portfolio returns. We then analyze a risk-averse investor's preferences by showing that he has indifference curves that imply a trade-off of expected return for variance.

[2]For example, expected utility can depend on the skewness (the third moment) of the return on wealth. The observation that some people purchase lottery tickets even though these investments have a negative expected rate of return suggests that their utility is enhanced by positive skewness. Alan Kraus and Robert Litzenberger (Kraus and Litzenberger 1976) developed a single-period portfolio selection and asset pricing model that extends Markowitz's analysis to consider investors who have a preference for skewness. Their results generalize Markowitz's model, but his fundamental insights are unchanged. For simplicity, this chapter focuses on the orginal Markowitz framework. Recent empirical work by Campbell Harvey and Akhtar Siddique (Harvey and Siddique 2000) examines the effect of skewness on asset pricing.

Subsequently, we show how a portfolio can be allocated among a given set of risky assets in a mean-variance efficient manner. We solve for the *efficient frontier*, defined as the set of portfolios that maximizes expected returns for a given variance of returns, and show that any two frontier portfolios can be combined to create a third. In addition, we show that a fundamental simplification to the investor's portfolio choice problem results when one of the assets included in the investor's choice set is a risk-free asset. The final section of this chapter applies mean-variance analysis to a problem of selecting securities to hedge the risk of commodity prices. This application is an example of how modern portfolio analysis has influenced the practice of risk management.

2.1 Assumptions on Preferences and Asset Returns

Suppose an expected-utility-maximizing individual invests her beginning-of-period wealth, W_0, in a particular portfolio of assets. Let $\widetilde{R}_p$ be the random return on this portfolio, so that the individual's end-of-period wealth is $\tilde{W} = W_0\widetilde{R}_p$.[3] Denote this individual's end-of-period utility by $U(\tilde{W})$. Given W_0, for notational simplicity we write $U(\tilde{W}) = U\left(W_0\widetilde{R}_p\right)$ as just $U(\widetilde{R}_p)$, because $\tilde{W}$ is completely determined by $\widetilde{R}_p$.

Let us express $U(\widetilde{R}_p)$ by expanding it in a Taylor series around the mean of $\widetilde{R}_p$, denoted as $E[\widetilde{R}_p]$. Let $U'(\cdot)$, $U''(\cdot)$, and $U^{(n)}(\cdot)$ denote the first, second, and n^{th} derivatives of the utility function:

[3] Thus, $\widetilde{R}_p$ is defined as one plus the rate of return on the portfolio.

$$\begin{aligned} U(\widetilde{R}_p) &= U\left(E[\widetilde{R}_p]\right) + \left(\widetilde{R}_p - E[\widetilde{R}_p]\right) U'\left(E[\widetilde{R}_p]\right) \\ &\quad + \tfrac{1}{2}\left(\widetilde{R}_p - E[\widetilde{R}_p]\right)^2 U''\left(E[\widetilde{R}_p]\right) + ... \\ &\quad + \tfrac{1}{n!}\left(\widetilde{R}_p - E[\widetilde{R}_p]\right)^n U^{(n)}\left(E[\widetilde{R}_p]\right) + ... \end{aligned} \tag{2.1}$$

Now let us investigate the conditions that would make this individual's expected utility depend only on the mean and variance of the portfolio return. We first analyze the restrictions on the form of utility, and then the restrictions on the distribution of asset returns, that would produce this result.

Note that if the utility function is quadratic, so that all derivatives of order 3 and higher are equal to zero ($U^{(n)} = 0, \forall\ n \geq 3$), then the individual's expected utility is

$$\begin{aligned} E\left[U(\widetilde{R}_p)\right] &= U\left(E[\widetilde{R}_p]\right) + \tfrac{1}{2}E\left[\left(\widetilde{R}_p - E[\widetilde{R}_p]\right)^2\right] U''\left(E[\widetilde{R}_p]\right) \\ &= U\left(E[\widetilde{R}_p]\right) + \tfrac{1}{2}V[\widetilde{R}_p]U''\left(E[\widetilde{R}_p]\right) \end{aligned} \tag{2.2}$$

where $V[\widetilde{R}_p]$ is the variance of the return on the portfolio.[4] Therefore, for any probability distribution of the portfolio return, $\widetilde{R}_p$, quadratic utility leads to expected utility that depends only on the mean and variance of $\widetilde{R}_p$.

Next, suppose that utility is not quadratic but any general increasing, concave form. Are there particular probability distributions for portfolio returns that make expected utility, again, depend only on the portfolio return's mean

[4] The expected value of the second term in the Taylor series, $E\left[\left(\widetilde{R}_p - E[\widetilde{R}_p]\right) U'\left(E[\widetilde{R}_p]\right)\right]$, equals zero.

and variance? Such distributions would need to be fully determined by their means and variances, that is, they must be two-parameter distributions whereby higher-order moments could be expressed in terms of the first two moments (mean and variance). Many distributions, such as the gamma, normal, and lognormal, satisfy this criterion. But in the context of an investor's portfolio selection problem, such distributions need to satisfy other reasonable conditions.

Since an individual is able to choose which assets to combine into a portfolio, all portfolios created from a combination of individual assets or other portfolios must have distributions that continue to be determined by their means and variances. In other words, we need a distribution such that if the individual assets' return distributions depend on just mean and variance, then the return on a linear combination (portfolio) of these assets has a distribution that depends on just the portfolio's mean and variance. Furthermore, the distribution should allow for a portfolio that possibly includes a risk-free (zero variance) asset, as well as assets that may be independently distributed. The only distributions that satisfy these "additivity," "possible risk-free asset," and "possible independent assets" restrictions is the stable family of distributions.[5] However, the only distribution within the stable family that has finite variance is the normal (Gaussian) distribution. Thus, since the multivariate normal distribution satisfies these portfolio conditions and has finite variance, it can be used to justify mean-variance analysis.

To verify that expected utility depends only on the portfolio return's mean and variance when this return is normally distributed, note that the third, fourth, and all higher central moments of the normal distribution are either zero or a function of the variance: $E\left[\left(\widetilde{R}_p - E[\widetilde{R}_p]\right)^n\right] = 0$ for n odd, and $E\left[\left(\widetilde{R}_p - E[\widetilde{R}_p]\right)^n\right] = \frac{n!}{(n/2)!}\left(\frac{1}{2}V[\widetilde{R}_p]\right)^{n/2}$ for n even. Therefore, in this case the individual's expected utility equals

[5] See (Chamberlain 1983) and (Liu 2004).

$$E\left[U(\widetilde{R}_p)\right] = U\left(E[\widetilde{R}_p]\right) + \tfrac{1}{2}V[\widetilde{R}_p]U''\left(E[\widetilde{R}_p]\right) + 0 + \frac{1}{8}\left(V[\widetilde{R}_p]\right)^2 U''''\left(E[\widetilde{R}_p]\right) + 0 + ... + \frac{1}{(n/2)!}\left(\frac{1}{2}V[\widetilde{R}_p]\right)^{n/2} U^{(n)}\left(E[\widetilde{R}_p]\right) + ... \quad (2.3)$$

which depends only on the mean and variance of the portfolio return.

In summary, restricting utility to be quadratic or restricting the distribution of asset returns to be normal allows us to write $E\left[U(\widetilde{R}_p)\right]$ as a function of only the mean, $E[\widetilde{R}_p]$, and the variance, $V[\widetilde{R}_p]$, of the portfolio return. Are either of these assumptions realistic? If not, it may be unjustified to suppose that only the first two moments of the portfolio return distribution matter to the individual investor.

The assumption of quadratic utility clearly is problematic. As mentioned earlier, quadratic utility displays negative marginal utility for levels of wealth greater than the "bliss point," and it has the unattractive characteristic of increasing absolute risk aversion. There are also difficulties with the assumption of normally distributed asset returns. When asset returns measured over any finite time period are normally distributed, there exists the possibility that their end-of-period values could be negative since realizations from the normal distribution have no lower (or upper) bound. This is an unrealistic description of returns for many assets such as stocks and bonds because, being limited-liability assets, their minimum value is nonnegative.[6]

As will be demonstrated in Chapter 12, the assumption of normal returns can be modified if we generalize the model to have multiple periods and assume that asset rates of return follow continuous-time stochastic processes. In that context, one can assume that assets' rates of return are *instantaneously* normally

[6] A related problem is that many standard utility functions, such as constant relative risk aversion, are not defined for negative values of portfolio wealth.

distributed, which implies that if their means and variances are constant over infinitesimal intervals, then over any finite interval asset values are lognormally distributed. This turns out to be a better way of modeling limited-liability assets because the lognormal distribution bounds these assets' values to be no less than zero. When we later study continuous-time, multiperiod models, we shall see that the results derived here assuming a single-period, discrete-time model continue to hold, under particular conditions, in the more realistic multi-period context. Moreover, in more complex multiperiod models that permit assets to have time-varying return distributions, we will show that optimal portfolio choices are straightforward generalizations of the mean-variance results derived in this chapter.

2.2 Investor Indifference Relations

Therefore, let us proceed by assuming that the individual's utility function, U, is a general concave utility function and that individual asset returns are normally distributed. Hence, a portfolio of these assets has a return $\tilde{R}_p$ that is normally distributed with probability density function $f(R; \bar{R}_p, \sigma_p^2)$, where we use the shorthand notation $\bar{R}_p \equiv E[\tilde{R}_p]$ and $\sigma_p^2 \equiv V[\tilde{R}_p]$. In this section we analyze an investor's "tastes," that is, the investor's risk-expected return preferences when utility depends on the mean (expected return) and variance (risk) of the return on wealth. The following section analyzes investment "technologies" represented by the combinations of portfolio risk and expected return that can be created from different portfolios of individual assets. Historically, mean-variance analysis has been illustrated graphically, and we will follow that convention while also providing analytic results.

Note that an investor's expected utility can then be written as

$$E\left[U\left(\widetilde{R}_p\right)\right] = \int_{-\infty}^{\infty} U(R) f(R; \bar{R}_p, \sigma_p^2) dR \tag{2.4}$$

To gain insight regarding this investor's preferences over portfolio risk and expected return, we wish to determine the characteristics of this individual's indifference curves in portfolio mean-variance space. An indifference curve represents the combinations of portfolio mean and variance that would give the individual the same level of expected utility.[7] To understand this relation, let us begin by defining $\widetilde{x} \equiv \frac{\tilde{R}_p - \bar{R}_p}{\sigma_p}$. Then

$$E\left[U\left(\widetilde{R}_p\right)\right] = \int_{-\infty}^{\infty} U(\bar{R}_p + x\sigma_p) n(x) dx \tag{2.5}$$

where $n(x) \equiv f(x; 0, 1)$ is the standardized normal probability density function, that is, the normal density having a zero mean and unit variance. Now consider how expected utility varies with changes in the mean and variance of the return on wealth. Taking the partial derivative with respect to $\bar{R}_p$:

$$\frac{\partial E\left[U\left(\widetilde{R}_p\right)\right]}{\partial \bar{R}_p} = \int_{-\infty}^{\infty} U' n(x) dx > 0 \tag{2.6}$$

since U' is always greater than zero. Next, take the partial derivative of equation (2.5) with respect to σ_p^2:

$$\frac{\partial E\left[U\left(\widetilde{R}_p\right)\right]}{\partial \sigma_p^2} = \frac{1}{2\sigma_p} \frac{\partial E\left[U\left(\widetilde{R}_p\right)\right]}{\partial \sigma_p} = \frac{1}{2\sigma_p} \int_{-\infty}^{\infty} U' x n(x) dx \tag{2.7}$$

While U' is always positive, x ranges between $-\infty$ and $+\infty$. Because x has a

[7] Indifference curves are used in microeconomics to analyze an individual's choice of consuming different quantities of multiple goods. For example, if utility, $u(x, y)$, derives from consuming two goods, with x being the quantity of good X consumed and y being the quantity of good Y consumed, then an indifference curve is the locus of points in X, Y space that gives a constant level of utility; that is, combinations of goods X and Y for which the individual would be indifferent between consuming. Mathematically, these combinations are represented as the points (x, y) such that $u(x, y) = \overline{U}$, a constant. In this section, we employ a similar concept but where expected utility depends on the mean and variance of the return on wealth.

standard normal distribution, which is symmetric, for each positive realization there is a corresponding negative realization with the same probability density. For example, take the positive and negative pair $+x_i$ and $-x_i$. Then $n(+x_i) = n(-x_i)$. Comparing the integrand of equation (2.7) for equal absolute realizations of x, we can show

$$\begin{aligned} & U'(\bar{R}_p + x_i\sigma_p)x_i n(x_i) + U'(\bar{R}_p - x_i\sigma_p)(-x_i)n(-x_i) \qquad (2.8) \\ = \; & U'(\bar{R}_p + x_i\sigma_p)x_i n(x_i) - U'(\bar{R}_p - x_i\sigma_p)x_i n(x_i) \\ = \; & x_i n(x_i)\left[U'(\bar{R}_p + x_i\sigma_p) - U'(\bar{R}_p - x_i\sigma_p)\right] < 0 \end{aligned}$$

because

$$U'(\bar{R}_p + x_i\sigma_p) < U'(\bar{R}_p - x_i\sigma_p) \qquad (2.9)$$

due to the assumed concavity of U; that is, the individual is risk averse so that $U'' < 0$. Thus, comparing $U'x_i n(x_i)$ for each positive and negative pair, we conclude that

$$\frac{\partial E\left[U\left(\widetilde{R}_p\right)\right]}{\partial \sigma_p^2} = \frac{1}{2\sigma_p}\int_{-\infty}^{\infty} U'xn(x)dx < 0 \qquad (2.10)$$

which is the intuitive result that higher portfolio variance, without higher portfolio expected return, reduces a risk-averse individual's expected utility.

Finally, an indifference curve is the combinations of portfolio mean and variance that leaves expected utility unchanged. In other words, it is combinations of $\left(\bar{R}_p, \sigma_p^2\right)$ that satisfy the equation $E\left[U\left(\widetilde{R}_p\right)\right] = \overline{U}$, a constant. Higher levels of $\overline{U}$ denote different indifference curves providing a greater level of utility. If we totally differentiate this equation, we obtain

$$dE\left[U\left(\widetilde{R}_p\right)\right] = \frac{\partial E\left[U\left(\widetilde{R}_p\right)\right]}{\partial \sigma_p^2} d\sigma_p^2 + \frac{\partial E\left[U\left(\widetilde{R}_p\right)\right]}{\partial \bar{R}_p} d\bar{R}_p = 0 \tag{2.11}$$

which, based on our previous results, tells us that each indifference curve is positively sloped in $\left(\bar{R}_p, \sigma_p^2\right)$ space:

$$\frac{d\bar{R}_p}{d\sigma_p^2} = -\frac{\partial E\left[U\left(\widetilde{R}_p\right)\right]}{\partial \sigma_p^2} \Big/ \frac{\partial E\left[U\left(\widetilde{R}_p\right)\right]}{\partial \bar{R}_p} > 0 \tag{2.12}$$

Thus, the indifference curve's slope in (2.12) quantifies the extent to which the individual requires a higher portfolio mean for accepting a higher portfolio variance.

Indifference curves are typically drawn in mean-standard deviation space, rather than mean - variance space, because standard deviations of returns are in the same unit of measurement as returns or interest rates (rather than squared returns). Figure 2.1 illustrates such a graph, where the arrow indicates an increase in the utility level, $\overline{U}$.[8] It is left as an end-of-chapter exercise to show that the curves are convex due to the assumed concavity of the utility function.

Having analyzed an investor's preferences over different combinations of portfolio means and standard deviations (or variances), let us consider next what portfolio means and standard deviations are possible given the available distributions of returns for individual assets.

2.3 The Efficient Frontier

The individual's optimal choice of portfolio mean and variance is determined by the point where one of these indifference curves is tangent to the set of means

[8] Clearly, these indifference curves cannot "cross" (intersect), because we showed that utility is always increasing in expected portfolio return for a given level of portfolio standard deviation.

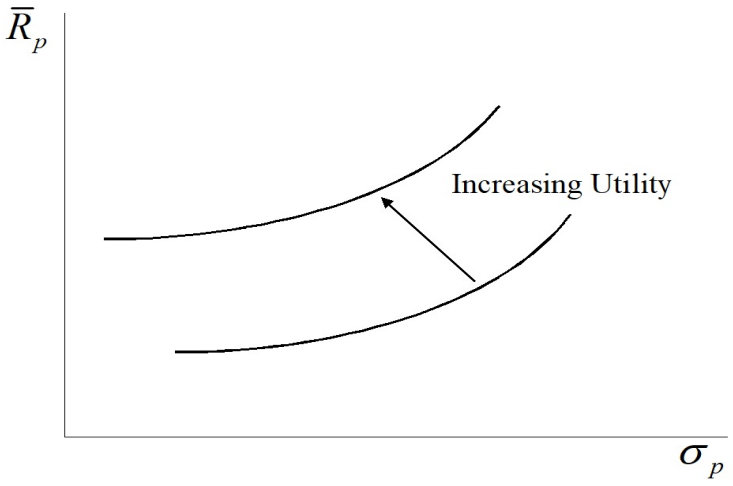


Figure 2.1: Indifference Curves

and standard deviations for all feasible portfolios, what we might describe as the "risk versus expected return investment opportunity set." This set represents all possible ways of combining various *individual* assets to generate alternative combinations of *portfolio* mean and variance (or standard deviation). This set includes inefficient portfolios (those in the interior of the opportunity set) as well as efficient portfolios (those on the "frontier" of the set). Efficient portfolios are those that make best use of the benefits of diversification. As we shall later prove, efficient portfolios have the attractive characteristic that any two efficient portfolios can be used to create any other efficient portfolio.

2.3.1 A Simple Example

To illustrate the effects of diversification, consider the following simple example. Suppose there are two assets, assets A and B, that have expected returns $\bar{R}_A$ and $\bar{R}_B$ and variances of σ_A^2 and σ_B^2, respectively. Further, the correlation

between their returns is given by ρ. Let us assume that $\bar{R}_A < \bar{R}_B$ but $\sigma_A^2 < \sigma_B^2$. Now form a portfolio with a proportion ω invested in asset A and a proportion $1-\omega$ invested in asset B.[9] The expected return on this portfolio is

$$\overline{R}_p = \omega \bar{R}_A + (1-\omega)\bar{R}_B \tag{2.13}$$

The expected return of a portfolio is a simple weighted average of the expected returns of the individual financial assets. Expected returns are not fundamentally transformed by combining individual assets into a portfolio. The standard deviation of the return on the portfolio is

$$\sigma_p = \left[\omega^2\sigma_A^2 + 2\omega(1-\omega)\sigma_A\sigma_B\rho + (1-\omega)^2\sigma_B^2\right]^{\frac{1}{2}} \tag{2.14}$$

In general, portfolio risk, as measured by the portfolio's return standard deviation, is a nonlinear function of the individual assets' variances and covariances. Thus, risk is altered in a relatively complex way when individual assets are combined in a portfolio.

Let us consider some special cases regarding the correlation between the two assets. Suppose $\rho = 1$, so that the two assets are perfectly positively correlated. Then the portfolio standard deviation equals

$$\begin{aligned} \sigma_p &= \left[\omega^2\sigma_A^2 + 2\omega(1-\omega)\sigma_A\sigma_B + (1-\omega)^2\sigma_B^2\right]^{\frac{1}{2}} \\ &= |\omega\sigma_A + (1-\omega)\sigma_B| \end{aligned} \tag{2.15}$$

which is a simple weighted average of the individual assets' standard deviations. Solving (2.15) for asset A's portfolio proportion gives $\omega = (\sigma_B \pm \sigma_p)/(\sigma_B - \sigma_A)$.

[9] It is assumed that ω can be any real number. A $\omega < 0$ indicates a short position in asset A, while $\omega > 1$ indicates a short position in asset B.

Then, by substituting for ω in (2.13), we obtain

$$\begin{aligned} \bar{R}_p &= \bar{R}_B + \left[\frac{\pm\sigma_p - \sigma_B}{\sigma_B - \sigma_A}\right](\bar{R}_B - \bar{R}_A) \\ &= \frac{\sigma_B\bar{R}_A - \sigma_A\bar{R}_B}{\sigma_B - \sigma_A} \pm \frac{\bar{R}_B - \bar{R}_A}{\sigma_B - \sigma_A}\sigma_p \end{aligned} \tag{2.16}$$

Thus, the relationship between portfolio risk and expected return are two straight lines in σ_p, $\bar{R}_p$ space. They have the same intercept of $\left(\sigma_B\bar{R}_A - \sigma_A\bar{R}_B\right)/(\sigma_B - \sigma_A)$ and have slopes of the same magnitude but opposite signs. The positively sloped line goes through the points $(\sigma_A, \bar{R}_A)$ and $(\sigma_B, \bar{R}_B)$ when $\omega = 1$ and $\omega = 0$, respectively. When $\omega = \sigma_B/(\sigma_B - \sigma_A) > 1$, indicating a short position in asset B, we see from (2.15) that all portfolio risk is eliminated ($\sigma_p = 0$). Figure 2.2 provides a graphical illustration of these relationships.

Next, suppose $\rho = -1$, so that the assets are perfectly negatively correlated. Then

$$\begin{aligned} \sigma_p &= \left[(\omega\sigma_A - (1-\omega)\sigma_B)^2\right]^{\frac{1}{2}} \\ &= |\omega\sigma_A - (1-\omega)\sigma_B| \end{aligned} \tag{2.17}$$

In a manner similar to the previous case, we can show that

$$\bar{R}_p = \frac{\sigma_A\bar{R}_B + \sigma_B\bar{R}_A}{\sigma_A + \sigma_B} \pm \frac{\bar{R}_B - \bar{R}_A}{\sigma_A + \sigma_B}\sigma_p \tag{2.18}$$

which, again, represents two straight lines in σ_p, $\bar{R}_p$ space. The intercept at $\sigma_p = 0$ is given by $\omega = \sigma_B/(\sigma_A + \sigma_B)$, so that all portfolio risk is eliminated with positive amounts invested in each asset. Furthermore, the negatively sloped line goes through the point $(\sigma_A, \bar{R}_A)$ when $\omega = 1$, while the positively sloped

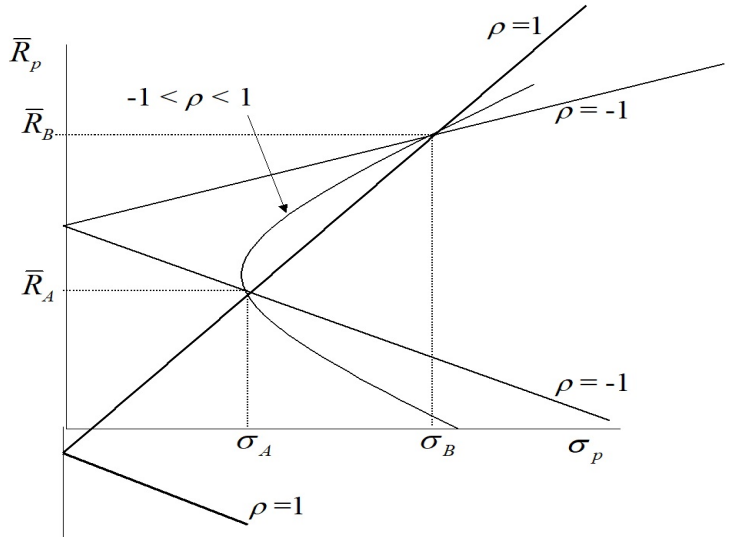


Figure 2.2: Efficient Frontier with Two Risky Assets

line goes through the point $(\sigma_B, \bar{R}_B)$ when $\omega = 0$. Figure 2.2 summarizes these risk-expected return constraints.

For either the $\rho = 1$ or $\rho = -1$ case, an investor would always choose a portfolio represented by the positively sloped lines because they give the highest average portfolio return for any given level of portfolio risk. These lines represent the so-called *efficient portfolio frontier*. The exact portfolio chosen by the individual would be where her indifference curve is tangent to the frontier.

When correlation between the assets is imperfect $(-1 < \rho < 1)$, the relationship between portfolio expected return and standard deviation is not linear but, as illustrated in Figure 2.2, is hyperbolic. In this case, it is no longer possible to create a riskless portfolio, so that the portfolio having minimum standard deviation is one where $\sigma_p > 0$. We now set out to prove these assertions for the general case of n assets.

2.3.2 Mathematics of the Efficient Frontier

Robert C. Merton (Merton 1972) provided an analytical solution to the following portfolio choice problem: Given the expected returns and the matrix of covariances of returns for n individual assets, find the set of portfolio weights that minimizes the variance of the portfolio for each feasible portfolio expected return. The locus of these points in portfolio expected return-variance space is the portfolio frontier. This section presents the derivation of Merton's solution. We begin by specifying the problem's notation and assumptions.

Let $\bar{R} = (\bar{R}_1 \bar{R}_2 ... \bar{R}_n)'$ be an $n \times 1$ vector of the expected returns of the n assets. Also let V be the $n \times n$ covariance matrix of the returns on the n assets. V is assumed to be of full rank.[10] Since it is a covariance matrix, it is also symmetric and positive definite. Next, let $\omega = (\omega_1 \ \omega_2 ... \omega_n)'$ be an $n \times 1$ vector of portfolio proportions, such that ω_i is the proportion of total portfolio wealth invested in the i^{th} asset. It follows that the expected return on the portfolio is given by

$$\bar{R}_p = \omega' \bar{R} \tag{2.19}$$

and the variance of the portfolio return is given by

$$\sigma_p^2 = \omega' V \omega \tag{2.20}$$

The constraint that the portfolio proportions must sum to 1 can be written as $\omega' e = 1$ where e is defined to be an $n \times 1$ vector of ones.

The problem of finding the portfolio frontier now can be stated as a quadratic optimization exercise: minimize the portfolio's variance subject to the con-

[10] This implies that there are no redundant assets among the n assets. An asset would be redundant if its return was an exact linear combination of the the returns on other assets. If such an asset exists, it can be ignored, since its availability does not affect the efficient portfolio frontier.

straints that the portfolio's expected return equals $\overline{R}_p$ and the portfolio's weights sum to one.[11]

$$\min_{\omega} \tfrac{1}{2}\omega' V \omega + \lambda \left[\overline{R}_p - \omega' \bar{R}\right] + \gamma[1 - \omega' e] \tag{2.21}$$

The first-order conditions with respect to ω and the two Lagrange multipliers, λ and γ, are

$$V\omega - \lambda \bar{R} - \gamma e = 0 \tag{2.22}$$

$$\overline{R}_p - \omega' \bar{R} = 0 \tag{2.23}$$

$$1 - \omega' e = 0 \tag{2.24}$$

Solving (2.22), the optimal portfolio weights satisfy

$$\omega^* = \lambda V^{-1} \bar{R} + \gamma V^{-1} e \tag{2.25}$$

Pre-multiplying equation (2.25) by $\bar{R}'$, we have

$$\overline{R}_p = \bar{R}' \omega^* = \lambda \bar{R}' V^{-1} \bar{R} + \gamma \bar{R}' V^{-1} e \tag{2.26}$$

Pre-multiplying equation (2.25) by e', we have

$$1 = e' \omega^* = \lambda e' V^{-1} \bar{R} + \gamma e' V^{-1} e \tag{2.27}$$

Equations (2.26) and (2.27) are two linear equations in two unknowns, λ and γ.

[11] In (2.21), the problem actually minimizes one-half the portfolio variance to avoid carrying an extra "2" in the first order condition (2.22). The solution is the same as minimizing the total variance and only changes the scale of the Lagrange multipliers.

The solution is

$$\lambda = \frac{\delta \overline{R}_p - \alpha}{\varsigma\delta - \alpha^2} \tag{2.28}$$

$$\gamma = \frac{\varsigma - \alpha \overline{R}_p}{\varsigma\delta - \alpha^2} \tag{2.29}$$

where $\alpha \equiv \bar{R}' V^{-1} e = e' V^{-1} \bar{R}$, $\varsigma \equiv \bar{R}' V^{-1} \bar{R}$, and $\delta \equiv e' V^{-1} e$ are scalars. Note that the denominators of λ and γ, given by $\varsigma\delta - \alpha^2$, are guaranteed to be positive when V is of full rank.[12] Substituting for λ and γ in equation (2.25), we have

$$\omega^* = \frac{\delta \overline{R}_p - \alpha}{\varsigma\delta - \alpha^2} V^{-1} \bar{R} + \frac{\varsigma - \alpha \overline{R}_p}{\varsigma\delta - \alpha^2} V^{-1} e \tag{2.30}$$

Collecting terms in $\bar{R}_p$ gives

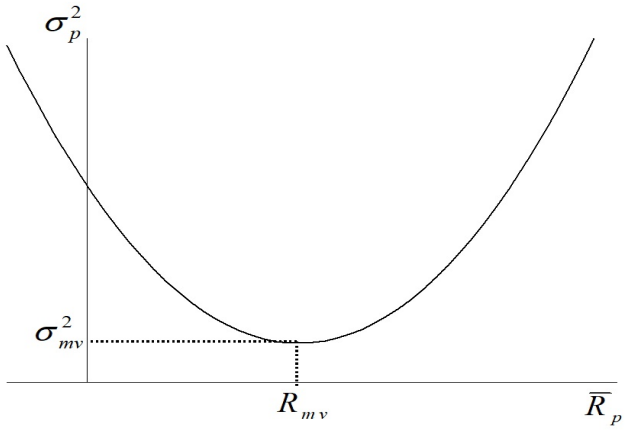
$$\omega^* = a + b\overline{R}_p \tag{2.31}$$

where $a \equiv \dfrac{\varsigma V^{-1} e - \alpha V^{-1} \bar{R}}{\varsigma\delta - \alpha^2}$ and $b \equiv \dfrac{\delta V^{-1} \bar{R} - \alpha V^{-1} e}{\varsigma\delta - \alpha^2}$. Equation (2.31) is both a necessary and sufficient condition for a frontier portfolio. Given $\overline{R}_p$, a portfolio must have weights satisfying (2.31) to minimize its return variance.

Having found the optimal portfolio weights for a given $\overline{R}_p$, the variance of the frontier portfolio is

$$\begin{aligned} \sigma_p^2 &= \omega^{*\prime} V \omega^* = (a + b\overline{R}_p)' V (a + b\overline{R}_p) \\ &= \frac{\delta \overline{R}_p^2 - 2\alpha \overline{R}_p + \varsigma}{\varsigma\delta - \alpha^2} \\ &= \frac{1}{\delta} + \frac{\delta \left(\overline{R}_p - \frac{\alpha}{\delta}\right)^2}{\varsigma\delta - \alpha^2} \end{aligned} \tag{2.32}$$

[12] To see this, note that since V is positive definite, so is V^{-1}. Therefore, the quadratic form $\left(\alpha\overline{R} - \varsigma e\right) V^{-1} \left(\alpha\overline{R} - \varsigma e\right) = \alpha^2\varsigma - 2\alpha^2\varsigma + \varsigma^2\delta = \varsigma\left(\varsigma\delta - \alpha^2\right)$ is positive. But since $\varsigma \equiv \overline{R} V^{-1} \overline{R}$ is a positive quadratic form, then $\left(\varsigma\delta - \alpha^2\right)$ must also be positive.



.pdf

Figure 2.3: Frontier Portfolios

where the second line in equation (2.32) results from substituting in the definitions of a and b and simplifying the resulting expression. Equation (2.32) is a parabola in σ_p^2, $\overline{R}_p$ space and is graphed in Figure 2.3. From the third line in equation (2.32), it is obvious that the unique minimum is at the point $\overline{R}_p = R_{mv} \equiv \frac{\alpha}{\delta}$, which corresponds to a global minimum variance of $\sigma_{mv}^2 \equiv \frac{1}{\delta}$. Substituting $\overline{R}_p = \frac{\alpha}{\delta}$ into equation (2.30) shows that this minimum variance portfolio has weights $\omega_{mv} = \frac{1}{\delta} V^{-1} e$.

Each point on the parabola in Figure 2.3 represents an investor's lowest possible portfolio variance, given some target level of expected return, $\overline{R}_p$. However, an investor whose utility is increasing in expected portfolio return and is decreasing in portfolio variance would never choose a portfolio having $\overline{R}_p < R_{mv}$, that is, points on the parabola to the left of R_{mv}. This is because the frontier portfolio's variance actually increases as the target expected return falls when $\overline{R}_p < R_{mv}$, making this target expected return region irrelevant to an optimizing investor. Hence, the *efficient* portfolio frontier is represented only by the region $\overline{R}_p \geq R_{mv}$.

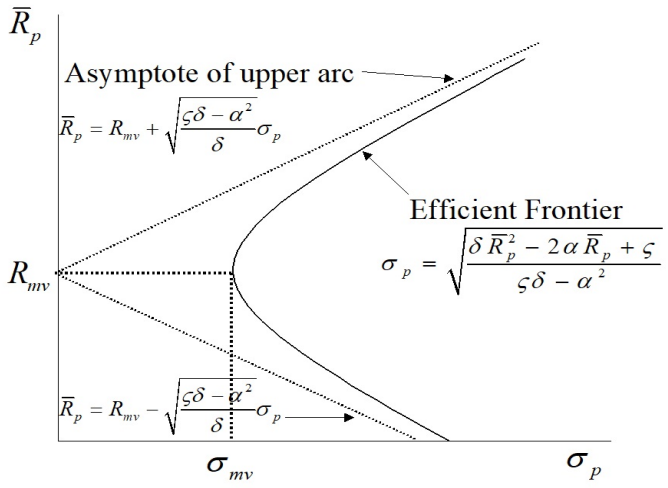


Figure 2.4: Efficient Frontier

Traditionally, portfolios satisfying (2.32) are graphed in σ_p, $\overline{R}_p$ space. Taking the square root of both sides of equation (2.32), σ_p becomes a hyperbolic function of $\overline{R}_p$. When this is graphed as in Figure 2.4 with $\overline{R}_p$ on the vertical axis and σ_p on the horizontal one, only the upper arc of the hyperbola is relevant because, as just stated, investors would not choose target levels of $\overline{R}_p < R_{mv}$. Differentiating (2.32), we can also see that the hyperbola's slope equals

$$\frac{\partial \overline{R}_p}{\partial \sigma_p} = \frac{\varsigma\delta - \alpha^2}{\delta\left(\overline{R}_p - \frac{\alpha}{\delta}\right)}\sigma_p \tag{2.33}$$

The upper arc asymptotes to the straight line $\overline{R}_p = R_{mv} + \sqrt{\frac{\varsigma\delta-\alpha^2}{\delta}}\sigma_p$, while the lower arc, representing inefficient frontier portfolios, asymptotes to the straight line $\overline{R}_p = R_{mv} - \sqrt{\frac{\varsigma\delta-\alpha^2}{\delta}}\sigma_p$.[13]

[13] To see that the slope of the hyperbola asymptotes to a magnitude of $\sqrt{(\varsigma\delta - \alpha^2)/\delta}$, use (2.32) to substitute for $\left(\overline{R}_p - \frac{\alpha}{\delta}\right)$ in (2.33) to obtain $\partial\overline{R}_p/\partial\sigma_p = \pm\sqrt{(\varsigma\delta - \alpha^2)}/\sqrt{\delta - 1/\sigma_p^2}$. Taking the limit of this expression as $\sigma_p \to \infty$ gives the desired result.

2.3.3 Portfolio Separation

We now state and prove a fundamental result:

> Every portfolio on the mean-variance frontier can be replicated by a combination of any two frontier portfolios; and an individual will be indifferent between choosing among the n financial assets, or choosing a combination of just two frontier portfolios.

This remarkable finding has an immediate practical implication. If all investors have the same beliefs regarding the distribution of asset returns, namely, returns are distributed $N\left(\overline{R}, V\right)$ and, therefore, the frontier is (2.32), then they can form their individually preferred frontier portfolios by trading in as little as two frontier portfolios. For example, if a security market offered two mutual funds, each invested in a different frontier portfolio, any mean-variance investor could replicate his optimal portfolio by appropriately dividing his wealth between only these two mutual funds.[14]

The proof of this separation result is as follows. Let $\bar{R}_{1p}$ and $\bar{R}_{2p}$ be the expected returns on any two distinct frontier portfolios. Let $\bar{R}_{3p}$ be the expected return on a third frontier portfolio. Now consider investing a proportion of wealth, x, in the first frontier portfolio and the remainder, $(1-x)$, in the second frontier portfolio. Clearly, a value for x can be found that makes the expected return on this "composite" portfolio equal to that of the third frontier portfolio:[15]

$$\bar{R}_{3p} = x\bar{R}_{1p} + (1-x)\bar{R}_{2p} \tag{2.34}$$

[14] To form his preferred frontier portfolio, an investor may require a short position in one of the frontier mutual funds. Since short positions are not possible with typical open-ended mutual funds, the better analogy would be that these funds are exchange-traded funds (ETFs) which do permit short positions.

[15] x may be any positive or negative number.

In addition, because portfolios 1 and 2 are frontier portfolios, we can write their portfolio proportions as a linear function of their expected returns. Specifically, we have $\omega^1 = a + b\bar{R}_{1p}$ and $\omega^2 = a + b\bar{R}_{2p}$ where ω^i is the $n \times 1$ vector of optimal portfolio weights for frontier portfolio i. Now create a new portfolio with an $n \times 1$ vector of portfolio weights given by $x\omega^1 + (1-x)\omega^2$. The portfolio proportions of this new portfolio can be written as

$$\begin{aligned} x\omega^1 + (1-x)\omega^2 &= x(a + b\bar{R}_{1p}) + (1-x)(a + b\bar{R}_{2p}) \\ &= a + b(x\bar{R}_{1p} + (1-x)\bar{R}_{2p}) \\ &= a + b\bar{R}_{3p} = \omega^3 \end{aligned} \tag{2.35}$$

where, in the last line of (2.35), we have substituted in equation (2.34). Based on the portfolio weights of the composite portfolio, $x\omega^1 + (1-x)\omega^2$, equaling $a + b\bar{R}_{3p}$, which represents the portfolio weights of the third frontier portfolio, ω^3, this composite portfolio equals the third frontier portfolio. Hence, any given efficient portfolio can be replicated by two frontier portfolios.

Portfolios on the mean-variance frontier have an additional property that will prove useful to the next section's analysis of portfolio choice when a riskless asset exists and also to understanding equilibrium asset pricing in Chapter 3. Except for the global minimum variance portfolio, ω_{mv}, for each frontier portfolio one can find another frontier portfolio with which its returns have zero covariance. That is, one can find pairs of frontier portfolios whose returns are orthogonal. To show this, note that the covariance between two frontier portfolios, w^1 and

w^2, is

$$\begin{aligned} \omega^1 V \omega^2 &= (a + b\overline{R}_{1p})' V (a + b\overline{R}_{2p}) \\ &= \frac{1}{\delta} + \frac{\delta}{\varsigma\delta - \alpha^2}\left(\overline{R}_{1p} - \frac{\alpha}{\delta}\right)\left(\overline{R}_{2p} - \frac{\alpha}{\delta}\right) \end{aligned} \tag{2.36}$$

Setting this equal to zero and solving for $\overline{R}_{2p}$, the expected return on the portfolio that has zero covariance with portfolio ω^1 is

$$\begin{aligned} \overline{R}_{2p} &= \frac{\alpha}{\delta} - \frac{\varsigma\delta - \alpha^2}{\delta^2\left(\overline{R}_{1p} - \frac{\alpha}{\delta}\right)} \\ &= R_{mv} - \frac{\varsigma\delta - \alpha^2}{\delta^2\left(\overline{R}_{1p} - R_{mv}\right)} \end{aligned} \tag{2.37}$$

Note that if $\left(\overline{R}_{1p} - R_{mv}\right) > 0$ so that frontier portfolio ω^1 is efficient, then equation (2.37) indicates that $\overline{R}_{2p} < R_{mv}$, implying that frontier portfolio ω^2 must be inefficient. We can determine the relative locations of these zero covariance portfolios by noting that in σ_p, $\overline{R}_p$ space, a line tangent to the frontier at the point $\left(\sigma_{1p}, \overline{R}_{1p}\right)$ is of the form

$$\overline{R}_p = \overline{R}_0 + \frac{\partial \overline{R}_p}{\partial \sigma_p}\Big|_{\sigma_p = \sigma_{1p}} \sigma_p \tag{2.38}$$

where $\frac{\partial \overline{R}_p}{\partial \sigma_p}\Big|_{\sigma_p = \sigma_{1p}}$ denotes the slope of the hyperbola at point $\left(\sigma_{1p}, \overline{R}_{1p}\right)$ and $\overline{R}_0$ denotes the tangent line's intercept at $\sigma_p = 0$. Using (2.33) and (2.32), we can solve for $\overline{R}_0$ by evaluating (2.38) at the point $\left(\sigma_{1p}, \overline{R}_{1p}\right)$:

$$\begin{aligned} \overline{R}_0 &= \overline{R}_{1p} - \frac{\partial \overline{R}_p}{\partial \sigma_p}\Big|_{\sigma_p = \sigma_{1p}} \sigma_{1p} = \overline{R}_{1p} - \frac{\varsigma\delta - \alpha^2}{\delta\left(\overline{R}_{1p} - \frac{\alpha}{\delta}\right)}\sigma_{1p}\sigma_{1p} \\ &= \overline{R}_{1p} - \frac{\varsigma\delta - \alpha^2}{\delta\left(\overline{R}_{1p} - \frac{\alpha}{\delta}\right)}\left[\frac{1}{\delta} + \frac{\delta\left(\overline{R}_{1p} - \frac{\alpha}{\delta}\right)^2}{\varsigma\delta - \alpha^2}\right] \\ &= \frac{\alpha}{\delta} - \frac{\varsigma\delta - \alpha^2}{\delta^2\left(\overline{R}_{1p} - \frac{\alpha}{\delta}\right)} \\ &= \overline{R}_{2p} \end{aligned} \tag{2.39}$$

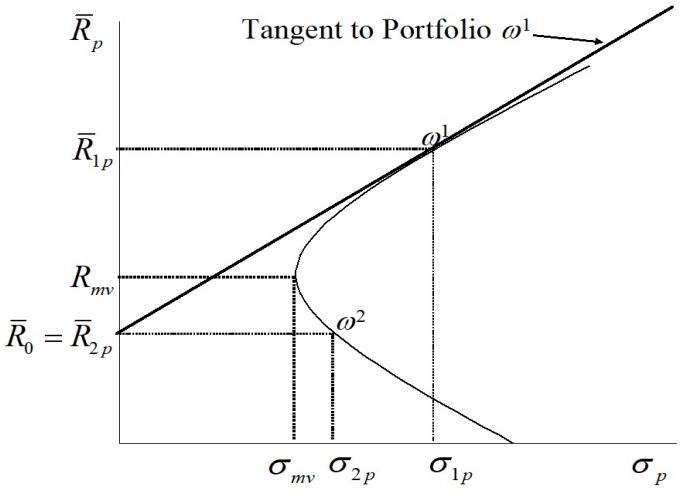


Figure 2.5: Frontier Portfolios with Zero Covariance

Hence, as shown in Figure 2.5, the intercept of the line tangent to frontier portfolio ω^1 equals the expected return of its zero-covariance counterpart, frontier portfolio ω^2.

2.4 The Efficient Frontier with a Riskless Asset

Thus far, we have assumed that investors can hold only risky assets. An implication of our analysis was that while all investors would choose efficient portfolios of risky assets, these portfolios would differ based on the particular investor's level of risk aversion. However, as we shall now see, introducing a riskless asset can simplify the investor's portfolio choice problem. This augmented portfolio choice problem, whose solution was first derived by James Tobin (Tobin 1958),

is one that we now consider.[16]

Assume that there is a riskless asset with return R_f. Let ω continue to be the $n \times 1$ vector of portfolio proportions invested in the risky assets. Now, however, the constraint $\omega' e = 1$ does not apply, because $1 - \omega' e$ is the portfolio proportion invested in the riskless asset. We can impose the restriction that the portfolio weights for all $n+1$ assets sum to one by writing the expected return on the portfolio as

$$\bar{R}_p = R_f + \omega'(\bar{R} - R_f e) \tag{2.40}$$

The variance of the return on the portfolio continues to be given by $\omega' V \omega$. Thus, the individual's optimization problem is changed to:

$$\min_{\omega} \tfrac{1}{2}\omega' V \omega + \lambda \left\{ \overline{R}_p - \left[R_f + \omega'(\bar{R} - R_f e) \right] \right\} \tag{2.41}$$

In a manner similar to the previous derivation, the first order conditions lead to the solution

$$\omega^* = \lambda V^{-1}(\bar{R} - R_f e) \tag{2.42}$$

where $\lambda \equiv \dfrac{\overline{R}_p - R_f}{\left(\bar{R} - R_f e\right)' V^{-1}(\bar{R} - R_f e)} = \dfrac{\overline{R}_p - R_f}{\varsigma - 2\alpha R_f + \delta R_f^2}$. Since V^{-1} is positive definite, λ is non-negative when $\overline{R}_p \geq R_f$, the region of the efficient frontier where investors' expected portfolio return is at least as great as the risk-free return. Given (2.42), the amount optimally invested in the riskless asset is determined by $1 - e' w^*$. Note that since λ is linear in $\overline{R}_p$, so is ω^*, similar to the previous case of no riskless asset. The variance of the portfolio now takes the form

[16] Tobin's work on portfolio selection was one of his contributions cited by the selection committee that awarded him the Nobel prize in economics in 1981.

$$\sigma_p^2 = \omega^{*\prime} V \omega^* = \frac{(\overline{R}_p - R_f)^2}{\varsigma - 2\alpha R_f + \delta R_f^2} \tag{2.43}$$

Taking the square root of each side of (2.43) and rearranging:

$$\overline{R}_p = R_f \pm \left(\varsigma - 2\alpha R_f + \delta R_f^2\right)^{\frac{1}{2}} \sigma_p \tag{2.44}$$

which indicates that the frontier is *linear* in σ_p, $\overline{R}_p$ space. Corresponding to the hyperbola for the no-riskless-asset case, the frontier when a riskless asset is included becomes two straight lines, each with an intercept of R_f but one having a positive slope of $\left(\varsigma - 2\alpha R_f + \delta R_f^2\right)^{\frac{1}{2}}$, the other having a negative slope of $-\left(\varsigma - 2\alpha R_f + \delta R_f^2\right)^{\frac{1}{2}}$. Of course, only the positively sloped line is the efficient portion of the frontier.

Since ω^* is linear in $\overline{R}_p$, the previous section's separation result continues to hold: any portfolio on the frontier can be replicated by two other frontier portfolios. However, when $R_f \neq R_{mv} \equiv \frac{\alpha}{\delta}$ holds, an even stronger separation principle obtains.[17] In this case, any portfolio on the linear efficient frontier can be replicated by two particular portfolios: one portfolio that is located on the "risky asset only" frontier and another portfolio that holds only the riskless asset.

Let us start by proving this result for the situation where $R_f < R_{mv}$. We assert that the efficient frontier given by the line $\overline{R}_p = R_f + \left(\varsigma - 2\alpha R_f + \delta R_f^2\right)^{\frac{1}{2}} \sigma_p$ can be replicated by a portfolio consisting of only the riskless asset and a portfolio on the risky-asset-only frontier that is determined by a straight line tangent to this frontier whose intercept is R_f. This situation is illustrated in Figure 2.6 where ω^A denotes the portfolio of risky assets determined by the tangent line

[17] We continue to let R_{mv} denote the expected return on the minimum variance portfolio that holds only risky assets. Of course, with a riskless asset, the minimum variance portfolio would be one that holds only the riskless asset.

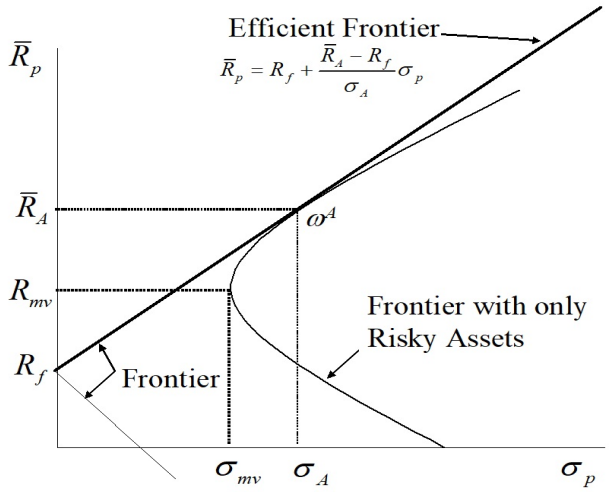


Figure 2.6: Efficient Frontier with a Riskless Asset

having intercept R_f. If we can show that the slope of this tangent line equals $\left(\varsigma - 2\alpha R_f + \delta R_f^2\right)^{\frac{1}{2}}$, then our assertion is proved.[18] Let $\overline{R}_A$ and σ_A be the expected return and standard deviation of return, respectively, on this tangency portfolio. Then the results of (2.37) and (2.39) allow us to write the slope of the tangent as

$$\begin{aligned} \frac{\overline{R}_A - R_f}{\sigma_A} &= \left[\frac{\alpha}{\delta} - \frac{\varsigma\delta - \alpha^2}{\delta^2\left(R_f - \frac{\alpha}{\delta}\right)} - R_f\right]/\sigma_A \\ &= \left[\frac{2\alpha R_f - \varsigma - \delta R_f^2}{\delta\left(R_f - \frac{\alpha}{\delta}\right)}\right]/\sigma_A \end{aligned} \tag{2.45}$$

Furthermore, we can use (2.32) and (2.37) to write

[18] Note that if a proportion x is invested in any risky asset portfolio having expected return and standard deviation of $\overline{R}_A$and σ_A, respectively, and a proportion $1 \quad x$ is invested in the riskless asset having certain return R_f, then the combined portfolio has an expected return and standard deviation of $\overline{R}_p = R_f + x\left(\overline{R}_A - R_f\right)$ and $\sigma_p = x\sigma_A$, respectively. When graphed in $\overline{R}_p, \sigma_p$ space, we can substitute for x to show that these combination portfolios are represented by the straight line $\overline{R}_p = R_f + \frac{R_A - R_f}{\sigma_A}\sigma_p$ whose intercept is R_f.

$$\begin{aligned} \sigma_A^2 &= \frac{1}{\delta} + \frac{\delta\left(\overline{R}_A - \frac{\alpha}{\delta}\right)^2}{\varsigma\delta - \alpha^2} \\ &= \frac{1}{\delta} + \frac{\varsigma\delta - \alpha^2}{\delta^3\left(R_f - \frac{\alpha}{\delta}\right)^2} \\ &= \frac{\delta R_f^2 - 2\alpha R_f + \varsigma}{\delta^2\left(R_f - \frac{\alpha}{\delta}\right)^2} \end{aligned} \tag{2.46}$$

Substituting the square root of (2.46) into (2.45) gives[19]

$$\begin{aligned} \frac{\overline{R}_A - R_f}{\sigma_A} &= \left[\frac{2\alpha R_f - \varsigma - \delta R_f^2}{\delta\left(R_f - \frac{\alpha}{\delta}\right)}\right] \frac{-\delta\left(R_f - \frac{\alpha}{\delta}\right)}{\left(\delta R_f^2 - 2\alpha R_f + \varsigma\right)^{\frac{1}{2}}} \\ &= \left(\delta R_f^2 - 2\alpha R_f + \varsigma\right)^{\frac{1}{2}} \end{aligned} \tag{2.47}$$

which is the desired result.

This result is an important simplification. If all investors agree on the distribution of asset returns (returns are distributed $N\left(\overline{R}, V\right)$), then they all consider the linear efficient frontier to be $\overline{R}_p = R_f + \left(\varsigma - 2\alpha R_f + \delta R_f^2\right)^{\frac{1}{2}} \sigma_p$ and all will choose to hold risky assets in the same relative proportions given by the tangency portfolio ω^A. Investors differ only in the amount of wealth they choose to allocate to this portfolio of risky assets versus the risk-free asset.

Along the efficient frontier depicted in Figure 2.7, the proportion of an investor's total wealth held in the tangency portfolio, $e'\omega^*$, increases as one moves to the right. At point $\left(\sigma_p, \overline{R}_p\right) = (0, R_f)$, $e'\omega^* = 0$ and all wealth is invested in the risk-free asset. In between points $(0, R_f)$ and $\left(\sigma_A, \overline{R}_A\right)$, which would be the case if, say, investor 1 had an indifference curve tangent to the efficient frontier at point $\left(\sigma_1, \overline{R}_{p1}\right)$, then $0 < e'\omega^* < 1$ and positive proportions of wealth

[19] Because it is assumed that $R_f < \frac{\alpha}{\delta}$, the square root of (2.46) has an opposite sign in order for σ_A to be positive.

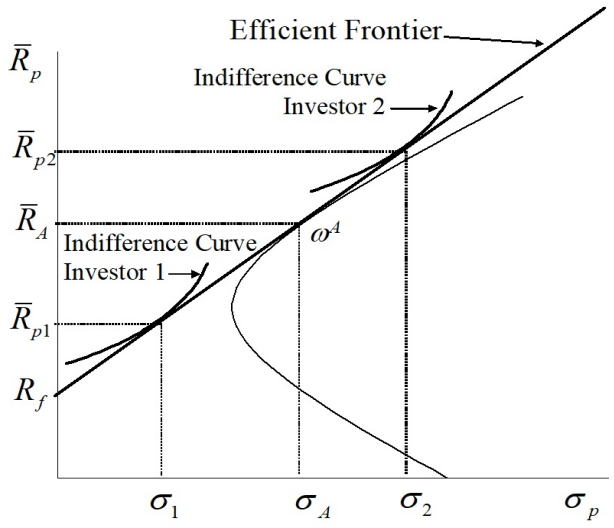


Figure 2.7: Investor Portfolio Choice

are invested in the risk-free asset and the tangency portfolio of risky assets. At point $\left(\sigma_A, \overline{R}_A\right)$, $e\acute{\omega}^* = 1$ and all wealth is invested in risky assets and none in the risk-free asset. Finally, to the right of this point, which would be the case if, say, investor 2 had an indifference curve tangent to the efficient frontier at point $\left(\sigma_2, \overline{R}_{p2}\right)$, then $e\acute{\omega}^* > 1$. This implies a negative proportion of wealth in the risk-free asset. The interpretation is that investor 2 borrows at the risk-free rate to invest more than 100 percent of her wealth in the tangency portfolio of risky assets. In practical terms, such an investor could be viewed as buying risky assets "on margin," that is, leveraging her asset purchases with borrowed money.

It will later be argued that $R_f < R_{mv}$, the situation depicted in Figures 2.6 and 2.7, is required for asset market equilibrium. However, we briefly describe the implications of other parametric cases. When $R_f > R_{mv}$, the efficient frontier of $\overline{R}_p = R_f + \left(\varsigma - 2\alpha R_f + \delta R_f^2\right)^{\frac{1}{2}} \sigma_p$ is always above the risky-

asset-only frontier. Along this efficient frontier, the investor short-sells the tangency portfolio of risky assets. This portfolio is located on the inefficient portion of the risky-asset-only frontier at the point where the line $\overline{R}_p = R_f - \left(\varsigma - 2\alpha R_f + \delta R_f^2\right)^{\frac{1}{2}} \sigma_p$ becomes tangent. The proceeds from this short-selling are then wholly invested in the risk-free asset. Lastly, when $R_f = R_{mv}$, the portfolio frontier is given by the asymptotes illustrated in Figure 2.4. It is straightforward to show that $\acute{e}\omega^* = 0$ for this case, so that total wealth is invested in the risk-free asset. However, the investor also holds a risky, but zero net wealth, position in risky assets. In other words, the proceeds from short-selling particular risky assets are used to finance long positions in other risky assets.

2.4.1 An Example with Negative Exponential Utility

To illustrate our results, let us specify a form for an individual's utility function. This enables us to determine the individual's preferred efficient portfolio, that is, the point of tangency between the individual's highest indifference curve and the efficient frontier. Given a specific utility function and normally distributed asset returns, we show how the individual's optimal portfolio weights can be derived directly by maximizing expected utility.

As before, let $\tilde{W}$ be the individual's end-of-period wealth and assume that she maximizes expected negative exponential utility:

$$U(\tilde{W}) = -e^{-b\tilde{W}} \tag{2.48}$$

where b is the individual's coefficient of absolute risk aversion. Now define $b_r \equiv bW_0$, which is the individual's coefficient of relative risk aversion at initial wealth W_0. Equation (2.48) can be rewritten:

$$U(\tilde{W}) = -e^{-b_r \tilde{W}/W_0} = -e^{-b_r \tilde{R}_p} \tag{2.49}$$

where $\tilde{R}_p$ is the total return (one plus the rate of return) on the portfolio.

In this problem, we assume that initial wealth can be invested in a riskless asset and n risky assets. As before, denote the return on the riskless asset as R_f and the returns on the n risky assets as the $n \times 1$ vector $\tilde{R}$. Also as before, let $\omega = (\omega_1 ... \omega_n)'$ be the vector of portfolio weights for the n risky assets. The risky assets' returns are assumed to have a joint normal distribution where $\bar{R}$ is the $n \times 1$ vector of expected returns on the n risky assets and V is the $n \times n$ covariance matrix of returns. Thus, the expected return on the portfolio can be written $\bar{R}_p \equiv R_f + \omega'(\bar{R} - R_f e)$ and the variance of the return on the portfolio is $\sigma_p^2 \equiv \omega' V \omega$.

Now recall the properties of the lognormal distribution. If $\tilde{x}$ is a normally distributed random variable, for example, $\tilde{x} \sim N(\mu, \sigma^2)$, then $\tilde{z} = e^{\tilde{x}}$ is lognormally distributed. The expected value of $\tilde{z}$ is

$$E[\tilde{z}] = e^{\mu + \frac{1}{2}\sigma^2} \tag{2.50}$$

From (2.49), we see that if $\tilde{R}_p = R_f + \omega'(\tilde{R} - R_f e)$ is normally distributed, then $U\left(\tilde{W}\right)$ is lognormally distributed. Using equation (2.50), we have

$$E\left[U\left(\widetilde{W}\right)\right] = -e^{-b_r\left[R_f + \omega'(\bar{R} - R_f e)\right] + \frac{1}{2} b_r^2 \omega' V \omega} \tag{2.51}$$

The individual chooses portfolio weights by maximizing expected utility:

$$\max_{\omega} E\left[U\left(\widetilde{W}\right)\right] = \max_{\omega} - e^{-b_r\left[R_f + \omega'(\bar{R} - R_f e)\right] + \frac{1}{2} b_r^2 \omega' V \omega} \tag{2.52}$$

Because the expected utility function is monotonic in its exponent, the maxi-

mization problem in (2.52) is equivalent to

$$\max_{\omega} \omega'(\bar{R} - R_f e) - \tfrac{1}{2} b_r \omega' V \omega \tag{2.53}$$

The n first-order conditions are

$$\bar{R} - R_f e - b_r V \omega = 0 \tag{2.54}$$

Solving for ω, we obtain

$$\omega^* = \frac{1}{b_r} V^{-1}(\bar{R} - R_f e) \tag{2.55}$$

Thus, we see that the individual's optimal portfolio choice depends on b_r, her coefficient of relative risk aversion, and the expected returns and covariances of the assets. Comparing (2.55) to (2.42), note that

$$\frac{1}{b_r} = \lambda \equiv \frac{\overline{R}_p - R_f}{\left(\bar{R} - R_f e\right)' V^{-1}(\bar{R} - R_f e)} \tag{2.56}$$

so that the greater the investor's relative risk aversion, b_r, the smaller is her target mean portfolio return, $\overline{R}_p$, and the smaller is the proportion of wealth invested in the risky assets. In fact, multiplying both sides of (2.55) by W_0, we see that the absolute amount of wealth invested in the risky assets is

$$W_0 \omega^* = \frac{1}{b} V^{-1}(\bar{R} - R_f e) \tag{2.57}$$

Therefore, the individual with constant absolute risk aversion, b, invests a fixed dollar amount in the risky assets, independent of her initial wealth. As wealth increases, each additional dollar is invested in the risk-free asset. Recall that this same result was derived at the end of Chapter 1 for the special case of a single risky asset.

As in this example, constant absolute risk aversion's property of making risky asset choice independent of wealth often allows for simple solutions to portfolio choice problems when asset returns are assumed to be normally distributed. However, the unrealistic implication that both wealthy and poor investors invest the same dollar amount in risky assets limits the empirical applications of using this form of utility. As we shall see in later chapters of this book, models where utility displays constant relative risk aversion are more typical.

2.5 An Application to Cross-Hedging

The following application of mean-variance analysis is based on Anderson and Danthine (Anderson and Danthine 1981). Consider a one-period model of an individual or institution that is required to buy or sell a commodity in the future and would like to hedge the risk of such a transaction by taking positions in futures (or other financial securities) markets. Assume that this financial operator is committed at the beginning of the period, date 0, to buy y units of a risky commodity at the end of the period, date 1, at the then prevailing spot price p_1. For example, a commitment to buy could arise if the commodity is a necessary input in the operator's production process.[20] Conversely, $y < 0$ represents a commitment to sell $-y$ units of a commodity, which could be due to the operator producing a commodity that is nonstorable.[21] What is important is that, as of date 0, y is deterministic, while p_1 is stochastic.

There are n financial securities (for example, futures contracts) in the economy. Denote the date 0 price of the i^{th} financial security as p^s_{i0}. Its date 1 price is p^s_{i1}, which is uncertain as of date 0. Let s_i denote the amount of the i^{th} security purchased at date 0. Thus, $s_i < 0$ indicates a short position in the

[20] An example of this case would be a utility that generates electricity from oil.

[21] For example, the operator could be a producer of an agricultural good, such as corn, wheat, or soybeans.

security.

Define the $n \times 1$ quantity and price vectors $s \equiv [s_1 \ldots s_n]'$, $p_0^s \equiv [p_{10}^s \ldots p_{n0}^s]'$, and $p_1^s \equiv [p_{11}^s \ldots p_{n1}^s]'$. Also define $p^s \equiv p_1^s - p_0^s$ as the $n \times 1$ vector of security price changes. This is the profit at date 1 from having taken unit long positions in each of the securities (futures contracts) at date 0, so that the operator's profit from its security position is $p^{s\prime}s$. Also define the first and second moments of the date 1 prices of the spot commodity and the financial securities: $E[p_1] = \bar{p}_1$, $Var[p_1] = \sigma_{00}$, $E[p_1^s] = \bar{p}_1^s$, $E[p^s] = \bar{p}^s$, $Cov[p_{i1}^s, p_{j1}^s] = \sigma_{ij}$, $Cov[p_1, p_{i1}^s] = \sigma_{0i}$, and the $(n+1) \times (n+1)$ covariance matrix of the spot commodity and financial securities is

$$\Sigma = \begin{bmatrix} \sigma_{00} & \Sigma_{01} \\ \Sigma_{01}' & \Sigma_{11} \end{bmatrix} \tag{2.58}$$

where Σ_{11} is an $n \times n$ matrix whose i, j^{th} element is σ_{ij}, and Σ_{01} is a $1 \times n$ vector whose i^{th} element is σ_{0i}.

For simplicity, let us assume that y is fixed and, therefore, is not a decision variable at date 0. Then the end-of-period profit (wealth) of the financial operator, W, is given by

$$W = p^{s\prime}s - p_1 y \tag{2.59}$$

What the operator must decide is the date 0 positions in the financial securities. We assume that the operator chooses s in order to maximize the following objective function that depends linearly on the mean and variance of profit:

$$\max_s E[W] - \tfrac{1}{2}\alpha Var[W] \tag{2.60}$$

As was shown in the previous section's equation (2.53), this objective func-

tion results from maximizing expected utility of wealth when portfolio returns are normally distributed and utility displays constant absolute risk aversion.[22] Substituting in for the operator's profit, we have

$$\max_{s} \bar{p}^{s\prime} s - \bar{p}_1 y - \tfrac{1}{2}\alpha \left[y^2 \sigma_{00} + s' \Sigma_{11} s - 2y \Sigma_{01} s \right] \tag{2.61}$$

The first-order conditions are

$$\bar{p}^s - \alpha \left[\Sigma_{11} s - y \Sigma_{01}' \right] = 0 \tag{2.62}$$

Thus, the optimal positions in financial securities are

$$\begin{aligned} s &= \frac{1}{\alpha} \Sigma_{11}^{-1} \bar{p}^s + y \Sigma_{11}^{-1} \Sigma_{01}' \\ &= \frac{1}{\alpha} \Sigma_{11}^{-1} \left(\bar{p}_1^s - p_0^s \right) + y \Sigma_{11}^{-1} \Sigma_{01}' \end{aligned} \tag{2.63}$$

Let us first consider the case of $y = 0$. This can be viewed as the situation faced by a pure *speculator*, by which we mean a trader who has no requirement to hedge. If $n = 1$ and $\bar{p}_1^s > p_0^s$, the speculator takes a long position in (purchases) the security, while if $\bar{p}_1^s < p_0^s$, the speculator takes a short position in (sells) the security. The magnitude of the position is tempered by the volatility of the security ($\Sigma_{11}^{-1} = 1/\sigma_{11}$), and the speculator's level of risk aversion, α. However, for the general case of $n > 1$, an expected price decline or rise is not sufficient to determine whether a speculator takes a long or short position in a particular security. All of the elements in Σ_{11}^{-1} need to be considered, since a position in

[22] Similar to the previous derivation, the objective function (2.60) can be derived from an expected utility function of the form $E[U(W)] = -\exp[-\alpha W]$ where α is the operator's coefficient of absolute risk aversion. Unlike the previous example, here the objective function is written in terms of total profit (wealth), not portfolio returns per unit wealth. Also, risky asset holdings, s, are in terms of absolute amounts purchased, not portfolio proportions. Hence, α is the coefficient of absolute risk aversion, not relative risk aversion.

a given security may have particular diversification benefits.

For the general case of $y \neq 0$, the situation faced by a *hedger*, the demand for financial securities is similar to that of a pure speculator in that it also depends on price expectations. In addition, there are hedging components to the demand for financial assets, call them s^h :

$$s^h \equiv y\Sigma_{11}^{-1}\Sigma_{01}' \tag{2.64}$$

This is the solution to the problem $\min_s \; Var(W)$. Thus, even for a hedger, it is never optimal to minimize volatility (risk) unless risk aversion is infinitely large. Even a risk-averse, expected-utility-maximizing hedger should behave somewhat like a speculator in that securities' expected returns matter. From definition (2.64), note that when $n = 1$ the pure hedging demand per unit of the commodity purchased, s^h/y, simplifies to[23]

$$\frac{s^h}{y} = \frac{Cov(p_1, p_1^s)}{Var(p_1^s)} \tag{2.65}$$

For the general case, $n > 1$, the elements of the vector $\Sigma_{11}^{-1}\Sigma_{01}'$ equal the coefficients $\beta_1, ..., \beta_n$ in the multiple regression model:

$$\Delta p_1 = \beta_0 + \beta_1 \Delta p_1^s + \beta_2 \Delta p_2^s + ... + \beta_n \Delta p_n^s + \varepsilon \tag{2.66}$$

where $\Delta p_1 \equiv p_1 - p_0$, $\Delta p_i^s \equiv p_{i1}^s - p_{i0}^s$, i = 1,...,n, and ε is a mean-zero error term. An implication of (2.66) is that an operator might estimate the *hedge ratios*, s^h/y, by performing a statistical regression using a historical times series of the $n \times 1$ vector of security price changes. In fact, this is a standard way that practitioners calculate hedge ratios.

[23] Note that if the correlation between the commodity price and the financial security return were equal to 1, so that a perfect hedge exists, then (2.65) becomes $s^h/y = \sqrt{\sigma_{00}}/\sqrt{\sigma_{11}}$; that is, the hedge ratio equals the ratio of the commodity price's standard deviation to that of the security price.

2.6 Summary

When the returns on individual assets are multivariate normally distributed, a risk-averse investor optimally chooses among a set of mean-variance efficient portfolios. Such portfolios make best use of the benefits of diversification by providing the highest mean portfolio return for a given portfolio variance. The particular efficient portfolio chosen by a given investor depends on her level of risk aversion. However, the ability to trade in only two efficient portfolios is sufficient to satisfy all investors, because any efficient portfolio can be created from any other two. When a riskless asset exists, the set of efficient portfolios has the characteristic that the portfolios' mean returns are linear in their portfolio variances. In such a case, a more risk-averse investor optimally holds a positive amount of the riskless asset and a positive amount of a particular risky-asset portfolio, while a less risk-averse investor optimally borrows at the riskless rate to purchase the same risky-asset portfolio in an amount exceeding his wealth.

This chapter provided insights on how individuals should optimally allocate their wealth among various assets. Taking the distribution of returns for all available assets as given, we determined any individual's portfolio demands for these assets. Having now derived a theory of investor asset demands, the next chapter will consider the equilibrium asset pricing implications of this investor behavior.

2.7 Exercises

1. Prove that the indifference curves graphed in Figure 2.1 are convex if the utility function is concave. Hint: suppose there are two portfolios, portfolios 1 and 2, that lie on the same indifference curve, where this

indifference curve has expected utility of $\overline{U}$. Let the mean returns on portfolios 1 and 2 be $\bar{R}_{1p}$ and $\bar{R}_{2p}$, respectively, and let the standard deviations of returns on portfolios 1 and 2 be σ_{1p} and σ_{2p}, respectively. Consider a third portfolio located in $(\overline{R}_p, \sigma_p)$ space that happens to be on a straight line between portfolios 1 and 2, that is, a portfolio having a mean and standard deviation satisfying $\bar{R}_{3p} = x\bar{R}_{1p} + (1-x)\bar{R}_{2p}$ and $\sigma_{3p} = x\sigma_{1p} + (1-x)\sigma_{2p}$ where $0 < x < 1$. Prove that the indifference curve is convex by showing that the expected utility of portfolio 3 exceeds $\overline{U}$. Do this by showing that the utility of portfolio 3 exceeds the convex combination of utilities for portfolios 1 and 2 for each standardized normal realization. Then integrate over all realizations to show this inequality holds for expected utilities.

2. Show that the covariance between the return on the minimum variance portfolio and the return on *any* other portfolio equals the variance of the return on the minimum variance portfolio. Hint: write down the variance of a portfolio that consists of a proportion x invested in the minimum variance portfolio and a proportion $(1-x)$ invested in any other portfolio. Then minimize the variance of this composite portfolio with respect to x.

3. Show how to derive the solution for the optimal portfolio weights for a frontier portfolio when there exists a riskless asset, that is, equation (2.42) given by $\omega^* = \lambda V^{-1}(\bar{R} - R_f e)$ where $\lambda \equiv \dfrac{\overline{R}_p - R_f}{\left(\bar{R} - R_f e\right)' V^{-1}(\bar{R} - R_f e)} = \dfrac{\overline{R}_p - R_f}{\varsigma - 2\alpha R_f + \delta R_f^2}$. The derivation is similar to the case with no riskless asset.

4. Show that when $R_f = R_{mv}$, the optimal portfolio involves $e'\omega^* = 0$.

5. Consider the mean-variance analysis covered in this chapter where there are n risky assets whose returns are jointly normally distributed. Assume

that investors differ with regard to their (concave) utility functions and their initial wealths. Also assume that investors can lend at the risk-free rate, $R_f < R_{mv}$, but investors are restricted from risk-free borrowing; that is, no risk-free borrowing is permitted.

a. Given this risk-free borrowing restriction, graphically show the efficient frontier for these investors in expected portfolio return-standard deviation space $\left(\overline{R}_p, \sigma_p\right)$.

b. Explain why only three portfolios are needed to construct this efficient frontier, and locate these three portfolios on your graph. (Note that these portfolios may not be unique.)

c. At least one of these portfolios will sometimes need to be sold short to generate the entire efficient frontier. Which portfolio(s) is it (label it on the graph) and in what range(s) of the efficient frontier will it be sold short? Explain.

6. Suppose there are n risky assets whose returns are multi-variate normally distributed. Denote their $n \times 1$ vector of expected returns as $\overline{R}$ and their $n \times n$ covariance matrix as V. Let there also be a riskless asset with return R_f. Let portfolio a be on the mean-variance efficient frontier and have an expected return and standard deviation of $\overline{R}_a$ and σ_a, respectively. Let portfolio b be any other (not necessarily efficient) portfolio having expected return and standard deviation $\overline{R}_b$ and σ_b, respectively. Show that the correlation between portfolios a and b equals portfolio b's Sharpe ratio divided by portfolio a's Sharpe ratio, where portfolio i's Sharpe ratio equals $\left(\bar{R}_i - R_f\right)/\sigma_i$. (Hint: write the correlation as $Cov\left(R_a, R_b\right)/\left(\sigma_a \sigma_b\right)$, and derive this covariance using the properties of portfolio efficiency.)

7. A corn grower has utility of wealth given by $U\left(W\right) = -e^{-aW}$ where a

> 0. This farmer's wealth depends on the total revenue from the sale of corn at harvest time. Total revenue is a random variable $\widetilde{s} = \widetilde{q}\widetilde{p}$, where $\widetilde{q}$ is the number of bushels of corn harvested and $\widetilde{p}$ is the spot price, net of harvesting costs, of a bushel of corn at harvest time. The farmer can enter into a corn futures contract having a current price of f_0 and a random price at harvest time of $\widetilde{f}$. If k is the number of short positions in this futures contract taken by the farmer, then the farmer's wealth at harvest time is given by $\widetilde{W} = \widetilde{s} - k\left(\widetilde{f} - f_0\right)$. If $\widetilde{s}\tilde{}\, N\left(\overline{s}, \sigma_s^2\right)$, $\widetilde{f}\tilde{}\, N\left(\overline{f}, \sigma_f^2\right)$, and $Cov\left(\widetilde{s}, \widetilde{f}\right) = \rho\sigma_s\sigma_f$, then solve for the optimal number of futures contract short positions, k, that the farmer should take.

8. Consider the standard Markowitz mean-variance portfolio choice problem where there are n risky assets and a risk-free asset. The risky assets' $n \times 1$ vector of returns, $\widetilde{R}$, has a multivariate normal distribution $N\left(\overline{R}, V\right)$, where $\overline{R}$ is the assets' $n \times 1$ vector of expected returns and V is a non-singular $n \times n$ covariance matrix. The risk-free asset's return is given by $R_f > 0$. As usual, assume no labor income so that the individual's end-of-period wealth depends only on her portfolio return; that is, $\widetilde{W} = W_0\widetilde{R}_p$, where the portfolio return is $\widetilde{R}_p = R_f + w'\left(\widetilde{R} - R_f e\right)$ where w is an $n \times 1$ vector of portfolio weights for the risky assets and e is an $n \times 1$ vector of 1s. Recall that we solved for the optimal portfolio weights, w^* for the case of an individual with expected utility displaying constant absolute risk aversion, $E\left[U\left(\widetilde{W}\right)\right] = E\left[-e^{-b\widetilde{W}}\right]$. Now, in this problem, consider the different case of an individual with expected utility displaying constant relative risk aversion, $E\left[U\left(\widetilde{W}\right)\right] = E\left[\frac{1}{\gamma}\widetilde{W}^{\gamma}\right]$ where $\gamma < 1$. What is w^* for this constant relative-risk-aversion case? Hint: recall the efficient frontier and consider the range of the probability distribution of the tangency portfolio. Also consider what would be the

individual's marginal utility should end-of-period wealth be nonpositive. This marginal utility will restrict the individual's optimal portfolio choice.

Chapter 3

CAPM, Arbitrage, and Linear Factor Models

In this chapter, we analyze the asset pricing implications of the previous chapter's mean-variance portfolio analysis. From one perspective, the Markowitz-Tobin portfolio selection rules form a normative theory instructing how an individual investor can best allocate wealth among various assets. However, these selection rules also could be interpreted as a positive or descriptive theory of how an investor actually behaves. If this latter view is taken, then a logical extension of portfolio selection theory is to consider the equilibrium asset pricing consequences of investors' individually rational actions. The portfolio choices of individual investors represent their particular demands for assets. By aggregating these investor demands and equating them to asset supplies, equilibrium asset prices can be determined. In this way, portfolio choice theory can provide a foundation for an asset pricing model. Indeed, such a model, the Capital Asset Pricing Model (CAPM), was derived at about the same time by

four individuals: Jack Treynor, William Sharpe, John Lintner, and Jan Mossin.[1] CAPM has influenced financial practice in highly diverse ways. It has provided foundations for capital budgeting rules, for the regulation of utilities' rates of return, for performance evaluation of money managers, and for the creation of indexed mutual funds.

This chapter starts by deriving the CAPM and studying its consequences for assets' rates of return. The notion that investors might require higher rates of return for some types of risks but not others is an important insight of CAPM and extends to other asset pricing models. CAPM predicts that assets' risk premia result from a single risk factor, the returns on the market portfolio of all risky assets which, in equilibrium, is a mean-variance efficient portfolio. However, it is not hard to imagine that a weakening of CAPM's restrictive assumptions could generate risk premia deriving from multiple factors. Hence, we then consider how assets' risk premia may be related when multiple risk factors generate assets' returns. We derive this relationship, not based on a model of investor preferences as was done in deriving CAPM, but based on the concept that competitive and efficient securities markets should not permit arbitrage.

As a prelude to considering a multifactor asset pricing model, we define and give examples of arbitrage. Arbitrage pricing is the primary technique for valuing one asset in terms of another. It is the basis of so-called relative pricing models, contingent claims models, or derivative pricing models. We look at some simple applications of arbitrage pricing and then study the multifactor Arbitrage Pricing Theory (APT) developed by Stephen Ross (Ross 1976). APT is the basis of the most popular empirical multifactor models of asset pricing.

[1] William Sharpe, a student of Harry Markowitz, shared the 1990 Nobel prize with Markowitz and Merton Miller. See (Treynor 1961), (Sharpe 1964), (Lintner 1965), and (Mossin 1966).

3.1 The Capital Asset Pricing Model

In Chapter 2, we proved that if investors maximize expected utility that depends only on the expected return and variance of end-of-period wealth, then no matter what their particular levels of risk aversion, they would be interested only in portfolios on the efficient frontier. This mean-variance efficient frontier was the solution to the problem of computing portfolio weights that would maximize a portfolio's expected return for a given portfolio standard deviation or, alternatively, minimizing a portfolio's standard deviation for a given expected portfolio return. The point on this efficient frontier ultimately selected by a given investor was that combination of expected portfolio return and portfolio standard deviation that maximized the particular investor's expected utility. For the case of n risky assets and a risk-free asset, the optimal portfolio weights for the n risky assets were shown to be

$$\omega^* = \lambda V^{-1}\left(\bar{R} - R_f e\right) \tag{3.1}$$

where $\lambda \equiv \dfrac{\overline{R}_p - R_f}{\varsigma - 2\alpha R_f + \delta R_f^2}$, $\alpha \equiv \bar{R}' V^{-1} e = e' V^{-1} \bar{R}$, $\varsigma \equiv \bar{R}' V^{-1} \bar{R}$, and $\delta \equiv e' V^{-1} e$. The amount invested in the risk-free asset is then $1 - e'\omega^*$. Since λ is a scalar quantity that is linear in $\overline{R}_p$, which is the individual investor's equilibrium portfolio expected return, the weights in equation (3.1) are also linear in $\overline{R}_p$. $\overline{R}_p$ is determined by where the particular investor's indifference curve is tangent to the efficient frontier. Thus, because differences in $\overline{R}_p$ just affect the scalar, λ, we see that all investors, no matter what their degree of risk aversion, choose to hold the risky assets in the same *relative* proportions.

Mathematically, we showed that the efficient frontier is given by

$$\sigma_p = \frac{\overline{R}_p - R_f}{\left(\varsigma - 2\alpha R_f + \delta R_f^2\right)^{\frac{1}{2}}} \tag{3.2}$$

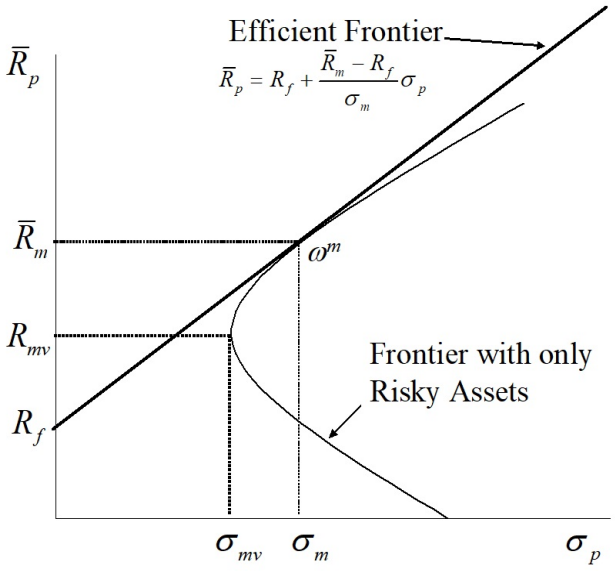


Figure 3.1: Capital Market Equilibrium

which, as illustrated in Figure 3.1, is linear when plotted in σ_p, $\overline{R}_p$ space.

3.1.1 Characteristics of the Tangency Portfolio

The efficient frontier, given by the line through R_f and ω^m, implies that investors optimally choose to hold combinations of the risk-free asset and the efficient frontier portfolio of risky assets having portfolio weights ω^m. We can easily solve for this unique "tangency" portfolio of risky assets since it is the point where an investor would have a zero position in the risk-free asset; that is, $e'\omega^* = 1$, or $\overline{R}_p = \bar{R}'\omega^*$. Pre-multiplying (3.1) by e', setting the result to 1, and solving for λ, we obtain $\lambda = m \equiv [\alpha - \delta R_f]^{-1}$, so that

$$\omega^m = mV^{-1}(\bar{R} - R_f e) \tag{3.3}$$

Let us now investigate the relationship between this tangency portfolio and individual assets. Consider the covariance between the tangency portfolio and the individual risky assets. Define σ_M as the $n \times 1$ vector of covariances of the tangency portfolio with each of the n risky assets. Then using (3.3) we see that

$$\sigma_M = Vw^m = m(\bar{R} - R_f e) \tag{3.4}$$

Note that the variance of the tangency portfolio is simply $\sigma_m^2 = \omega^{m\prime} V \omega^m$. Accordingly, if we then pre-multiply equation (3.4) by $\omega^{m\prime}$, we obtain

$$\begin{aligned} \sigma_m^2 &= \omega^{m\prime}\sigma_M = m\omega^{m\prime}(\bar{R} - R_f e) \\ &= m(\bar{R}_m - R_f) \end{aligned} \tag{3.5}$$

where $\bar{R}_m \equiv \omega^{m\prime}\bar{R}$ is the expected return on the tangency portfolio.[2] Rearranging (3.4) and substituting in for m from (3.5), we have

$$(\bar{R} - R_f e) = \frac{1}{m}\sigma_M = \frac{\sigma_M}{\sigma_m^2}(\bar{R}_m - R_f) = \beta(\bar{R}_m - R_f) \tag{3.6}$$

where $\beta \equiv \frac{\sigma_M}{\sigma_m^2}$ is the $n \times 1$ vector whose i^{th} element is $\frac{Cov(\tilde{R}_m, \tilde{R}_i)}{Var(\tilde{R}_m)}$. Equation (3.6) shows that a simple relationship links the excess expected return (expected return in excess of the risk-free rate) on the tangency portfolio, $(\bar{R}_m - R_f)$, to the excess expected returns on the individual risky assets, $(\bar{R} - R_f e)$.

[2] Note that the elements of ω^m sum to 1 since the tangency portfolio has zero weight in the risk-free asset.

3.1.2 Market Equilibrium

Now suppose that individual investors, each taking the set of individual assets' expected returns and covariances as fixed (exogenous), all choose mean-variance efficient portfolios. Thus, each investor decides to allocate his or her wealth between the risk-free asset and the unique tangency portfolio. Because individual investors demand the risky assets in the same relative proportions, we know that the *aggregate demands* for the risky assets will have the same relative proportions, namely, those of the tangency portfolio. Recall that our derivation of this result *does not* assume a "representative" investor in the sense of requiring all investors to have identical utility functions or beginning-of-period wealth. It *does* assume that investors have identical beliefs regarding the probability distribution of asset returns, that all risky assets can be traded, that there are no indivisibilities in asset holdings, and that there are no limits on borrowing or lending at the risk-free rate.

We can now define an equilibrium as a situation where asset returns are such that the investors' demands for the assets equal the assets' supplies. What determines the assets' supplies? One way to model asset supplies is to assume they are fixed. For example, the economy could be characterized by a fixed *quantity* of physical assets that produce random output at the end of the period. Such an economy is often referred to as an *endowment economy*, and we detail a model of this type in Chapter 6. In this case, equilibrium occurs by adjustment of the date 0 assets' prices so that investors' demands conform to the inelastic assets' supplies. The change in the assets' date 0 prices effectively adjusts the assets' return distributions to those which make the tangency portfolio and the net demand for the risk-free asset equal to the fixed supplies of these assets.

An alternative way to model asset supplies is to assume that the economy's asset return distributions are fixed but allow the quantities of these assets to be

elastically supplied. This type of economy is known as a *production economy*, and a model of it is presented in Chapter 13. Such a model assumes that there are n risky, constant-returns-to-scale "technologies." These technologies require date 0 investments of physical capital and produce end-of-period physical investment returns having a distribution with mean $\bar{R}$ and a covariance matrix of V at the end of the period. Also, there could be a risk-free technology that generates a one-period return on physical capital of R_f. In this case of a fixed return distribution, supplies of the assets adjust to the demands for the tangency portfolio and the risk-free asset determined by the technological return distribution.

As it turns out, how one models asset supplies does not affect the results that we now derive regarding the equilibrium relationship between asset returns. We simply note that the tangency portfolio having weights ω^m must be the equilibrium portfolio of risky assets supplied in the market. Thus, equation (3.6) can be interpreted as an equilibrium relationship between the excess expected return on any asset and the excess expected return on the *market* portfolio. In other words, in equilibrium, the tangency portfolio chosen by all investors must be the market portfolio of all risky assets. Moreover, as mentioned earlier, the only case for which investors have a long position in the tangency portfolio is $R_f < R_{mv}$. Hence, for asset markets to clear, that is, for the outstanding stocks of assets to be owned by investors, the situation depicted in Figure 3.1 can be the only equilibrium efficient frontier.[3]

The Capital Asset Pricing Model's prediction that the market portfolio is mean-variance efficient is an important solution to the practical problem of identifying a mean-variance efficient portfolio. As a theory, CAPM justifies the

[3]This presumes that the tangency portfolio is composed of long positions in the individual risky assets; that is, $\omega_i^m > 0$ for $i = 1, ..., n$. While our derivation has not restricted the sign of these portfolio weights, since assets must have nonnegative supplies, equilibrium market clearing implies that assets' prices or individuals' choice of technologies must adjust (effectively changing $\overline{R}$ and/or V) to make the portfolio demands for individual assets nonnegative.

practice of investing in a broad market portfolio of stocks and bonds. This insight has led to the growth of "indexed" mutual funds and exchange-traded funds (ETFs) that hold market-weighted portfolios of stocks and bonds.

Let's now look at some additional implications of CAPM when we consider realized, rather than expected, asset returns. Note that asset i's realized return, $\tilde{R}_i$, can be defined as $\bar{R}_i + \tilde{\nu}_i$, where $\tilde{\nu}_i$ is the unexpected component of the asset's return. Similarly, the realized return on the market portfolio, $\tilde{R}_m$, can be defined as $\bar{R}_m + \tilde{\nu}_m$, where $\tilde{\nu}_m$ is the unexpected part of the market portfolio's return. Substituting these into (3.6), we have

$$\begin{aligned} \tilde{R}_i &= R_f + \beta_i(\tilde{R}_m - \tilde{\nu}_m - R_f) + \tilde{\nu}_i \\ &= R_f + \beta_i(\tilde{R}_m - R_f) + \tilde{\nu}_i - \beta_i\tilde{\nu}_m \\ &= R_f + \beta_i(\tilde{R}_m - R_f) + \widetilde{\varepsilon}_i \end{aligned} \tag{3.7}$$

where $\widetilde{\varepsilon}_i \equiv \tilde{\nu}_i - \beta_i\tilde{\nu}_m$. Note that

$$\begin{aligned} Cov(\tilde{R}_m, \widetilde{\varepsilon}_i) &= Cov(\tilde{R}_m, \tilde{\nu}_i) - \beta_i Cov(\tilde{R}_m, \tilde{\nu}_m) \\ &= Cov(\tilde{R}_m, \tilde{R}_i) - \beta_i Cov(\tilde{R}_m, \tilde{R}_m) \\ &= \beta_i Var(\tilde{R}_m) - \beta_i Var(\tilde{R}_m) = 0 \end{aligned} \tag{3.8}$$

which, along with (3.7), implies that the total variance of risky asset i, σ_i^2, has two components:

$$\sigma_i^2 = \beta_i^2\sigma_m^2 + \sigma_{\varepsilon_i}^2 \tag{3.9}$$

where $\beta_i^2\sigma_m^2$ is proportional to the return variance of the market portfolios and $\sigma_{\varepsilon_i}^2$ is the variance of $\widetilde{\varepsilon}_i$, and it is orthogonal to the market portfolio's return.

Since equation (3.8) shows that $\widetilde{\varepsilon}_i$ is the part of the return on risky asset i that is uncorrelated with the return on the market portfolio, this implies that equation (3.7) represents a regression equation. In other words, an unbiased estimate of β_i can be obtained by running an Ordinary Least Squares regression of asset i's excess return on the market portfolio's excess return. The orthogonal, mean-zero residual, $\widetilde{\varepsilon}_i$, is sometimes referred to as idiosyncratic, unsystematic, or diversifiable risk. This is the particular asset's risk that is eliminated or diversified away when the asset is held in the market portfolio. Since this portion of the asset's risk can be eliminated by the individual who invests optimally, there is no "price" or "risk premium" attached to it in the sense that the asset's equilibrium expected return is not altered by it.

To make clear what risk is priced, let us denote the covariance between the return on the i^{th} asset and the return on the market portfolio as $\sigma_{Mi} = Cov(\tilde{R}_m, \widetilde{R}_i)$, which is the i^{th} element of σ_M. Also let ρ_{im} be the correlation between the return on the i^{th} asset and the return on the market portfolio. Then equation (3.6) can be rewritten as

$$\begin{aligned} \bar{R}_i - R_f &= \frac{\sigma_{Mi}}{\sigma_m}\frac{(\bar{R}_m - R_f)}{\sigma_m} \\ &= \rho_{im}\sigma_i\frac{(\bar{R}_m - R_f)}{\sigma_m} \\ &= \rho_{im}\sigma_i S_e \end{aligned} \tag{3.10}$$

where $S_e \equiv \frac{(\bar{R}_m - R_f)}{\sigma_m}$ is the equilibrium excess return on the market portfolio per unit of market risk and is known as the market Sharpe ratio, named after William Sharpe, one of the developers of the CAPM. S_e can be interpreted as the market price of systematic or nondiversifiable risk. It is also referred to as the slope of the *capital market line*, where the capital market line is defined as

the efficient frontier that connects the points R_f and ω^m in Figure 3.1. Now if we define ω_i^m as the weight of asset i in the market portfolio and V_i as the i^{th} row of covariance matrix V, then

$$\frac{\partial \sigma_m}{\partial \omega_i^m} = \frac{1}{2\sigma_m}\frac{\partial \sigma_m^2}{\partial \omega_i^m} = \frac{1}{2\sigma_m}\frac{\partial \omega^m V \omega^m}{\partial \omega_i^m} = \frac{1}{2\sigma_m} 2V_i\omega^m = \frac{1}{\sigma_m}\sum_{j=1}^{n}\omega_j^m\sigma_{ij} \qquad (3.11)$$

where σ_{ij} is the i,j^{th} element of V. Since $\tilde{R}_m = \sum_{j=1}^{n}\omega_j^m\tilde{R}_j$, then $Cov(\tilde{R}_i, \tilde{R}_m) = Cov(\tilde{R}_i, \sum_{j=1}^{n}\omega_j^m\tilde{R}_j) = \sum_{j=1}^{n}\omega_j^m\sigma_{ij}$. Hence, (3.11) can be rewritten as

$$\frac{\partial \sigma_m}{\partial \omega_i^m} = \frac{1}{\sigma_m}Cov(\tilde{R}_i, \tilde{R}_m) = \rho_{im}\sigma_i \qquad (3.12)$$

Thus, $\rho_{im}\sigma_i$ can be interpreted as the marginal increase in "market risk," σ_m, from a marginal increase of asset i in the market portfolio. In this sense, $\rho_{im}\sigma_i$ is the *quantity* of asset i's systematic or nondiversifiable risk. Equation (3.10) shows that this quantity of systematic risk, multiplied by the price of systematic risk, S_e, determines the asset's required excess expected return, or risk premium.

If a riskless asset does not exist so that all assets are risky, Fischer Black (Black 1972) showed that a similar asset pricing relationship exists. Here, we outline his *zero-beta CAPM*. Note that an implication of the portfolio separation result of section 2.3.3 is that since every frontier portfolio can be written as $\omega = a + b\bar{R}_p$, a linear combination of these frontier portfolios is also a frontier portfolio. Let W_i be the proportion of the economy's total wealth owned by investor i, and let ω^i be this investor's desired frontier portfolio so that $\omega^i = a + b\bar{R}_{ip}$. If there are a total of I investors, then the weights of the market portfolio are given by

$$\begin{aligned} \omega^m &= \sum_{i=1}^{I} W_i \omega^i = \sum_{i=1}^{I} W_i \left(a + b\bar{R}_{ip}\right) \\ &= a\sum_{i=1}^{I} W_i + b\sum_{i=1}^{I} W_i \bar{R}_{ip} = a + b\bar{R}_m \end{aligned} \tag{3.13}$$

where $\bar{R}_m \equiv \sum_{i=1}^{I} W_i \bar{R}_{ip}$ and where the last equality of (3.13) uses the fact that the sum of the proportions of total wealth must equal 1. Equation (3.13) shows that the market portfolio, the aggregation of all individual investors' portfolios, is a frontier portfolio. Its expected return, $\overline{R}_m$, is a weighted average of the expected returns of the individual investors' portfolios. Because each individual investor optimally chooses a portfolio on the efficient portion of the frontier (the upper arc in Figure 2.4), then the market portfolio, being a weighted average, is also on the efficient frontier.

Now, let us compute the covariance between the market portfolio and any arbitrary portfolio of risky assets, not necessarily a frontier portfolio. Let this arbitrary risky-asset portfolio have weights ω^0, a random return of $\widetilde{R}_{0p}$, and an expected return of $\overline{R}_{0p}$. Then

$$\begin{aligned} Cov\left(\widetilde{R}_m, \widetilde{R}_{0p}\right) &= \omega^{m\prime} V \omega^0 = \left(a + b\bar{R}_m\right)' V\omega^0 \\ &= \left(\frac{\varsigma V^{-1}e - \alpha V^{-1}\bar{R}}{\varsigma\delta - \alpha^2} + \frac{\delta V^{-1}\bar{R} - \alpha V^{-1}e}{\varsigma\delta - \alpha^2}\bar{R}_m\right)' V\omega^0 \\ &= \frac{\varsigma e' V^{-1} V\omega^0 - \alpha\bar{R}' V^{-1} V\omega^0}{\varsigma\delta - \alpha^2} \\ &\quad + \frac{\delta\bar{R}_m\bar{R}' V^{-1} V\omega^0 - \alpha\bar{R}_m e' V^{-1} V\omega^0}{\varsigma\delta - \alpha^2} \\ &= \frac{\varsigma - \alpha\overline{R}_{0p} + \delta\bar{R}_m\overline{R}_{0p} - \alpha\bar{R}_m}{\varsigma\delta - \alpha^2} \end{aligned} \tag{3.14}$$

Rearranging (3.14) gives

$$\overline{R}_{0p} = \frac{\alpha \bar{R}_m - \varsigma}{\delta \bar{R}_m - \alpha} + Cov\left(\widetilde{R}_m, \widetilde{R}_{0p}\right) \frac{\varsigma\delta - \alpha^2}{\delta \bar{R}_m - \alpha} \tag{3.15}$$

Rewriting the first term on the right-hand side of equation (3.15) and multiplying and dividing the second term by the definition of a frontier portfolio's variance given in Chapter 2's equation (2.32), equation (3.15) becomes

$$\begin{aligned} \overline{R}_{0p} &= \frac{\alpha}{\delta} - \frac{\varsigma\delta - \alpha^2}{\delta^2\left(\overline{R}_m - \frac{\alpha}{\delta}\right)} + \frac{Cov\left(\widetilde{R}_m, \widetilde{R}_{0p}\right)}{\sigma_m^2} \left(\frac{1}{\delta} + \frac{\delta\left(\overline{R}_m - \frac{\alpha}{\delta}\right)^2}{\varsigma\delta - \alpha^2}\right) \frac{\varsigma\delta - \alpha^2}{\delta \bar{R}_m - \alpha} \\ &= \frac{\alpha}{\delta} - \frac{\varsigma\delta - \alpha^2}{\delta^2\left(\overline{R}_m - \frac{\alpha}{\delta}\right)} + \frac{Cov\left(\widetilde{R}_m, \widetilde{R}_{0p}\right)}{\sigma_m^2} \left(\overline{R}_m - \frac{\alpha}{\delta} + \frac{\varsigma\delta - \alpha^2}{\delta^2\left(\overline{R}_m - \frac{\alpha}{\delta}\right)}\right) \end{aligned} \tag{3.16}$$

From equation (2.39), we recognize that the first two terms on the right-hand side of (3.16) equal the expected return on the portfolio that has zero covariance with the market portfolio, call it $\overline{R}_{zm}$. Thus, equation (3.16) can be written as

$$\begin{aligned} \overline{R}_{0p} &= \overline{R}_{zm} + \frac{Cov\left(\widetilde{R}_m, \widetilde{R}_{0p}\right)}{\sigma_m^2}\left(\overline{R}_m - \overline{R}_{zm}\right) \\ &= \overline{R}_{zm} + \beta_0\left(\overline{R}_m - \overline{R}_{zm}\right) \end{aligned} \tag{3.17}$$

Since the portfolio having weights ω^0 can be any risky-asset portfolio, it includes a portfolio that invests solely in a single asset.[4] In this light, β_0 becomes the covariance of the individual asset's return with that of the market portfolio, and the relationship in equation (3.17) is identical to the previous CAPM result in equation (3.10) except that $\overline{R}_{zm}$ replaces R_f. Hence, when a riskless asset does not exist, we measure an asset's excess returns relative to $\overline{R}_{zm}$, the expected return on a portfolio that has a zero beta.

[4] One of the elements of ω^0 would equal 1, while the rest would be zero.

Because the CAPM relationship in equations (3.10) or (3.17) implies that assets' expected returns differ only due to differences in their betas, it is considered a single "factor" model, this risk factor being the return on the market portfolio. Stephen Ross (Ross 1976) derived a similar multifactor relationship, but starting from a different set of assumptions and using a derivation based on the arbitrage principle. Frequently in this book, we will see that asset pricing implications can often be derived based on investor risk preferences, as was done in the CAPM when we assumed investors cared only about the mean and variance of their portfolio's return. However, another powerful technique for asset pricing is to rule out the existence of arbitrage. We now turn to this topic, first by discussing the nature of arbitrage.

3.2 Arbitrage

The notion of arbitrage is simple. It involves the possibility of getting something for nothing while having no possibility of loss. Specifically, consider constructing a portfolio involving both long and short positions in assets such that no initial wealth is required to form the portfolio.[5] If this zero-net-investment portfolio can sometimes produce a positive return but can never produce a negative return, then it represents an arbitrage: starting from zero wealth, a profit can sometimes be made but a loss can never occur. A special case of arbitrage is when this zero-net-investment portfolio produces a riskless return. If this certain return is positive (*negative*), an arbitrage is to buy (*sell*) the portfolio and reap a riskless profit, or "free lunch." Only if the return is zero would there be no arbitrage.

An arbitrage opportunity can also be defined in a slightly different context.

[5] Proceeds from short sales (or borrowing) are used to purchase (take long positions in) other assets.

If a portfolio that requires a nonzero initial net investment is created such that it earns a certain rate of return, then this rate of return must equal the current (competitive market) risk-free interest rate. Otherwise, there would also be an arbitrage opportunity. For example, if the portfolio required a positive initial investment but earned less than the risk-free rate, an arbitrage would be to (short-) sell the portfolio and invest the proceeds at the risk-free rate, thereby earning a riskless profit equal to the difference between the risk-free rate and the portfolio's certain (lower) rate of return.[6]

In efficient, competitive asset markets where arbitrage trades are feasible, it is reasonable to think that arbitrage opportunities are rare and fleeting. Should arbitrage temporarily exist, then trading by investors to earn this profit will tend to move asset prices in a direction that eliminates the arbitrage opportunity. For example, if a zero-net-investment portfolio produces a riskless positive return, as investors create (buy) this portfolio, the prices of the assets in the portfolio will be bid up. The cost of creating the portfolio will then exceed zero. The portfolio's cost will rise until it equals the present value of the portfolio's riskless return, thereby eliminating the arbitrage opportunity. Hence, for competitive asset markets where it is also feasible to execute arbitrage trades, it may be reasonable to assume that equilibrium asset prices reflect an absence of arbitrage opportunities. As will be shown, this assumption leads to a law of one price: if different assets produce exactly the same future payoffs, then the current prices of these assets must be the same. This simple result has powerful asset pricing implications.

[6] Arbitrage defined in this context is really equivalent to the previous definition of arbitrage. For example, if a portfolio requiring a positive initial investment produces a certain rate of return in excess of the riskless rate, then an investor should be able to borrow the initial funds needed to create this portfolio and pay an interest rate on this loan that equals the risk-free interest rate. That the investor should be able to borrow at the riskless interest rate can be seen from the fact that the portfolio produces a return that is always sufficient to repay the loan in full, making the borrowing risk-free. Hence, combining this initial borrowing with the nonzero portfolio investment results in an arbitrage opportunity that requires zero initial wealth.

However, as a word of caution, not all asset markets meet the conditions required to justify arbitrage pricing. For some markets, it may be impossible to execute pure arbitrage trades due to significant transactions costs and/or restrictions on short-selling or borrowing. In such cases of limited arbitrage, the law of one price can fail.[7] Alternative methods, such as those based on a model of investor preferences, are required to price assets.

3.2.1 Examples of Arbitrage Pricing

An early use of the arbitrage principle is the *covered interest parity* condition that links spot and forward foreign exchange markets to foreign and domestic money markets. To illustrate, let $F_{0\tau}$ be the current date 0 forward price for exchanging one unit of a foreign currency τ periods in the future. This forward price represents the dollar price to be paid τ periods in the future for delivery of one unit of foreign currency τ periods in the future. In contrast, let S_0 be the spot price of foreign exchange, that is, the current date 0 dollar price of one unit of foreign currency to be delivered immediately. Also let R_f be the per-period risk-free (money market) return for borrowing or lending in dollars over the period 0 to τ, and denote as R_f^* the per-period risk-free return for borrowing or lending in the foreign currency over the period 0 to τ.[8]

Now construct the following portfolio that requires zero net wealth. First, we sell forward (take a short forward position in) one unit of foreign exchange at price $F_{0\tau}$.[9] This contract means that we are committed to delivering one unit of foreign exchange at date τ in return for receiving $F_{0\tau}$ dollars at date τ. Second, let us also purchase the present value of one unit of foreign currency,

[7] Andrei Shleifer and Robert Vishny (Shleifer and Vishny 1997) discuss why the conditions needed to apply arbitrage pricing are not present in many asset markets.

[8] For example, if the foreign currency is the Japanese yen, R_f^* would be the per-period return for a yen-denominated risk-free investment or loan.

[9] Taking a long or short position in a forward contract requires zero initial wealth, as payment and delivery all occur at the future date τ.

$1/R_f^{*\tau}$, and invest it in a foreign bond yielding the per-period return, R_f^*. In terms of the domestic currency, this purchase costs $S_0/R_f^{*\tau}$, which we finance by borrowing dollars at the per-period return R_f.

What happens at date τ as a result of these trades? When date τ arrives, we know that our foreign currency investment yields $R_f^{*\tau}/R_f^{*\tau} = 1$ unit of the foreign currency. This is exactly what we need to satisfy our short position in the forward foreign exchange contract. For delivering this foreign currency, we receive $F_{0\tau}$ dollars. But we also now owe a sum of $R_f^\tau S_0/R_f^{*\tau}$ due to our dollar borrowing. Thus, our net proceeds at date τ are

$$F_{0\tau} - R_f^\tau S_0/R_f^{*\tau} \tag{3.18}$$

Note that these proceeds are nonrandom; that is, the amount is known at date 0 since it depends only on prices and riskless rates quoted at date 0. If this amount is positive, then we should indeed create this portfolio as it represents an arbitrage. If, instead, this amount is negative, then an arbitrage would be for us to sell this portfolio; that is, we reverse each trade just discussed (i.e., take a long forward position, and invest in the domestic currency financed by borrowing in foreign currency markets). Thus, the only instance in which arbitrage would not occur is if the net proceeds are zero, which implies

$$F_{0\tau} = S_0 R_f^\tau / R_f^{*\tau} \tag{3.19}$$

Equation (3.19) is referred to as the *covered interest parity* condition.

The forward exchange rate, $F_{0\tau}$, represents the dollar price for buying or selling a foreign currency at date τ, a future date when the foreign currency's dollar value is unknown. Though $F_{0\tau}$ is the price of a risky cashflow, it has been determined without knowledge of the utility functions of investors or their

expectations regarding the future value of the foreign currency. The reason for this simplification is due to the *law of one price*, which states that in the absence of arbitrage, equivalent assets (or contracts) must have the same price. A forward contract to purchase a unit of foreign currency can be replicated by buying, at the spot exchange rate S_0, a foreign currency investment paying the per-period, risk-free return R_f^* and financing this by borrowing at the dollar risk-free return R_f. In the absence of arbitrage, these two methods for obtaining foreign currency in the future must be valued the same. Given the spot exchange rate, S_0, and the foreign and domestic money market returns, R_f^* and R_f, the forward rate is pinned down. Thus, when applicable, pricing assets or contracts by ruling out arbitrage is attractive in that assumptions regarding investor preferences or beliefs are not required.

To motivate how arbitrage pricing might apply to a very simple version of the CAPM, suppose that there is a risk-free asset that returns R_f and multiple risky assets. However, assume that only a single source of (market) risk determines all risky-asset returns and that these returns can be expressed by the linear relationship

$$\widetilde{R}_i = a_i + b_i \tilde{f} \tag{3.20}$$

where $\widetilde{R}_i$ is the return on the i^{th} asset and $\tilde{f}$ is the single risk factor generating all asset returns, where it is assumed that $E[\tilde{f}] = 0$. a_i is asset i's expected return, that is, $E[\widetilde{R}_i] = a_i$. b_i is the sensitivity of asset i to the risk factor and can be viewed as asset i's beta coefficient. Note that this is a highly simplified example in that all risky assets are perfectly correlated with each other. Assets have no idiosyncratic risk (residual component $\widetilde{\varepsilon}_i$). A generalized model with idiosyncratic risk will be presented in the next section.

Now suppose that a portfolio of two assets is constructed, where a proportion

of wealth of ω is invested in asset i and the remaining proportion of $(1-\omega)$ is invested in asset j. This portfolio's return is given by

$$\begin{aligned} \widetilde{R}_p &= \omega a_i + (1-\omega)a_j + \omega b_i \tilde{f} + (1-\omega) b_j \tilde{f} \\ &= \omega(a_i - a_j) + a_j + [\omega(b_i - b_j) + b_j]\, \tilde{f} \end{aligned} \tag{3.21}$$

If the portfolio weights are chosen such that

$$\omega^* = \frac{b_j}{b_j - b_i} \tag{3.22}$$

then the uncertain (random) component of the portfolio's return is eliminated. The absence of arbitrage then requires that $R_p = R_f$, so that

$$R_p = \omega^*(a_i - a_j) + a_j = R_f \tag{3.23}$$

or

$$\frac{b_j(a_i - a_j)}{b_j - b_i} + a_j = R_f$$

which implies

$$\frac{a_i - R_f}{b_i} = \frac{a_j - R_f}{b_j} \equiv \lambda \tag{3.24}$$

This condition states that the expected return in excess of the risk-free rate, per unit of risk, must be equal for all assets, and we define this ratio as λ. λ is the risk premium per unit of the factor risk. The denominator, b_i, can be interpreted as asset i's quantity of risk from the single risk factor, while $a_i - R_f$ can be thought of as asset i's compensation or premium in terms of excess expected return given to investors for holding asset i. Thus, this no-arbitrage

condition is really a law of one price in that the *price of risk*, λ, which is the risk premium divided by the quantity of risk, must be the same for all assets.

Equation (3.24) is a fundamental relationship, and similar law-of-one-price conditions hold for virtually all asset pricing models. For example, we can rewrite the CAPM equation (3.10) as

$$\frac{\bar{R}_i - R_f}{\rho_{im}\sigma_i} = \frac{(\bar{R}_m - R_f)}{\sigma_m} \equiv S_e \tag{3.25}$$

so that the ratio of an asset's expected return premium, $\bar{R}_i - R_f$, to its quantity of market risk, $\rho_{im}\sigma_i$, is the same for all assets and equals the slope of the capital market line, S_e. We next turn to a generalization of the CAPM that derives from arbitrage pricing.

3.3 Linear Factor Models

The CAPM assumption that all assets can be held by all individual investors is clearly an oversimplification. Transactions costs and other trading "frictions" that arise from distortions such as capital controls and taxes might prevent individuals from holding a global portfolio of marketable assets. Furthermore, many assets simply are nonmarketable and cannot be traded.[10] The preeminent example of a nonmarketable asset is the value of an individual's future labor income, what economists refer to as the individual's *human capital*. Therefore, in addition to the risk from returns on a global portfolio of marketable assets, individuals are likely to face multiple sources of nondiversifiable risks. It is then not hard to imagine that, in equilibrium, assets' risk premia derive from

[10] Richard Roll (Roll 1977) has argued that CAPM is not a reasonable theory, because a true "market" portfolio consisting of all risky assets cannot be observed or owned by investors. Moreover, empirical tests of CAPM are infeasible because proxies for the market portfolio (such as the S&P 500 stock index) may not be mean-variance efficient, even if the true market portfolio is. Conversely, a proxy for the market portfolio could be mean-variance efficient even though the true market portfolio is not.

more than a single risk factor. Indeed, the CAPM's prediction that risk from a market portfolio is the only source of priced risk has not received strong empirical support.[11]

This is a motivation for the multifactor Arbitrage Pricing Theory (APT) model. APT assumes that an individual asset's return is driven by multiple risk factors and by an idiosyncratic component, though the theory is mute regarding the sources of these multiple risk factors. APT is a relative pricing model in the sense that it determines the risk premia on all assets relative to the risk premium for each of the factors and each asset's sensitivity to each factor.[12] It does not make assumptions regarding investor preferences but uses arbitrage pricing to restrict an asset's risk premium. The main assumptions of the model are that the returns on all assets are linearly related to a finite number of risk factors and that the number of assets in the economy is large relative to the number of factors. Let us now detail the model's assumptions.

Assume that there are k risk factors and n assets in the economy, where $n > k$. Let b_{iz} be the sensitivity of the i^{th} asset to the z^{th} risk factor, where $\tilde{f}_z$ is the random realization of risk factor z. Also let $\widetilde{\varepsilon}_i$ be the idiosyncratic risk component specific to asset i, which by definition is independent of the k risk factors, $\tilde{f}_1$,...,$\tilde{f}_k$, and the specific risk component of any other asset j, $\widetilde{\varepsilon}_j$. $\widetilde{\varepsilon}_i$ must be independent of the risk factors or else it would affect all assets, thus not being truly a specific source of risk to just asset i. If a_i is the expected return on asset i, then the return-generating process for asset i is given by the linear factor model

[11] Ravi Jagannathan and Ellen McGrattan (Jagannathan and McGrattan 1995) review the empirical evidence for CAPM.

[12] This is not much different from the CAPM. CAPM determined each asset's risk premium based on the single-factor market risk premium, $\overline{R}_m - R_f$, and the asset's sensitivity to this single factor, β_i. The only difference is that CAPM provides somewhat more guidance as to the identity of the risk factor, namely, the return on a market portfolio of all assets.

$$\widetilde{R}_i = a_i + \sum_{z=1}^{k} b_{iz}\tilde{f}_z + \widetilde{\varepsilon}_i \tag{3.26}$$

where $E[\widetilde{\varepsilon}_i] = E\left[\tilde{f}_z\right] = E\left[\widetilde{\varepsilon}_i\tilde{f}_z\right] = 0$, and $E[\widetilde{\varepsilon}_i\widetilde{\varepsilon}_j] = 0$ for $i \neq j$. For simplicity, we also assume that $E\left[\tilde{f}_z\tilde{f}_x\right] = 0$ for $z \neq x$; that is, the risk factors are mutually independent. In addition, let us further assume that the risk factors are normalized to have a variance equal to one, so that $E\left[\tilde{f}_z^2\right] = 1$. As it turns out, these last two assumptions are not important, as a linear transformation of correlated risk factors can allow them to be redefined as independent, unit-variance risk factors.[13]

A final assumption is that the idiosyncratic risk (variance) for each asset is finite; that is,

$$E\left[\widetilde{\varepsilon}_i^2\right] \equiv s_i^2 < S^2 \tag{3.27}$$

where S^2 is some finite number. Under these assumptions, note that $Cov\left(\widetilde{R}_i, \tilde{f}_z\right) = Cov\left(b_{iz}\tilde{f}_z, \tilde{f}_z\right) = b_{iz}Cov\left(\tilde{f}_z, \tilde{f}_z\right) = b_{iz}$. Thus, b_{iz} is the covariance between the return on asset i and factor z.

In the simple example of the previous section, assets had no idiosyncratic risk, and their expected returns could be determined by ruling out a simple arbitrage. This was because a hedge portfolio, consisting of appropriate combinations of different assets, could be created that had a riskless return. Now, however, when each asset's return contains an idiosyncratic risk component, it is not possible to create a hedge portfolio having a purely riskless return. Instead, we will argue that if the number of assets is large, a portfolio can be constructed that has "close" to a riskless return, because the idiosyncratic components of

[13] For example, suppose $\widetilde{g}$ is a $k \times 1$ vector of mean-zero, correlated risk factors with $k \times k$ covariance matrix $E[\widetilde{g}\widetilde{g}'] = \Omega$. Then create a transformed $k \times 1$ vector of risk factors given by $\widetilde{f} = \sqrt{\Omega^{-1}}\widetilde{g}$. The covariance matrix of these transformed risk factors is $E\left[\widetilde{f}\widetilde{f}'\right] = \sqrt{\Omega^{-1}}E[\widetilde{g}\widetilde{g}']\sqrt{\Omega^{-1}} = I_k$ where I_k is a $k \times k$ identity matrix.

assets' returns are diversifiable. While ruling out pure arbitrage opportunities is not sufficient to constrain assets' expected returns, we can use the notion of *asymptotic arbitrage* to argue that assets' expected returns will be "close" to the relationship that would result if they had no idiosyncratic risk. So let us now state what we mean by an asymptotic arbitrage opportunity.[14]

Definition: Let a portfolio containing n assets be described by the vector of investment amounts in each of the n assets, $W^n \equiv [W_1^n \; W_2^n \; ... \; W_n^n]'$. Thus, W_i^n is the amount invested in asset i when there are n total assets in the economy. Consider a sequence of these portfolios where n is increasing, $n = 2, 3, \ldots$. Let σ_{ij} be the covariance between the returns on assets i and j. Then an asymptotic arbitrage exists if the following conditions hold:

(A) The portfolio requires zero net investment:

$$\sum_{i=1}^{n} W_i^n = 0$$

(B) The portfolio return becomes certain as n gets large:

$$\lim_{n\to\infty} \sum_{i=1}^{n} \sum_{j=1}^{n} W_i^n W_j^n \sigma_{ij} = 0$$

(C) The portfolio's expected return is always bounded above zero

$$\sum_{i=1}^{n} W_i^n a_i \geq \delta > 0$$

We can now state the Arbitrage Pricing Theorem (APT):

Theorem: If no asymptotic arbitrage opportunities exist, then the expected return of asset i, $i = 1, ..., n$, is described by the following linear relation:

[14] This proof of Arbitrage Pricing Theory based on the concept of asymptotic arbitrage is due to Gur Huberman (Huberman 1982).

$$a_i = \lambda_0 + \sum_{z=1}^{k} b_{iz}\lambda_z + \nu_i \qquad (*)$$

where λ_0 is a constant, λ_z is the risk premium for risk factor $\widetilde{f}_z$, $z = 1, ..., k$, and the expected return deviations, ν_i, satisfy

$$\sum_{i=1}^{n} \nu_i = 0 \qquad (i)$$

$$\sum_{i=1}^{n} b_{iz}\nu_i = 0, \quad z = 1, ..., k \qquad (ii)$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \nu_i^2 = 0 \qquad (iii)$$

Note that condition (iii) says that the average squared error (deviation) from the pricing rule $(*)$ goes to zero as n becomes large. Thus, as the number of assets increases relative to the risk factors, expected returns will, on average, become closely approximated by the relation $a_i = \lambda_0 + \sum_{z=1}^{k} b_{iz}\lambda_z$. Also note that if the economy contains a risk-free asset (implying $b_{iz} = 0, \forall\, z$), the risk-free return will be approximated by λ_0.

Proof: For a given number of assets, $n > k$, think of running a cross-sectional regression of the a_i's on the b_{iz}'s. More precisely, project the dependent variable vector $a = [a_1\ a_2 \dots a_n]'$ on the k explanatory variable vectors $b_z = [b_{1z}\ b_{2z} \dots b_{nz}]'$, $z = 1, ..., k$. Define ν_i as the regression residual for observation i, $i = 1, ..., n$. Denote λ_0 as the regression intercept and λ_z, $z = 1, ..., k$, as the estimated coefficient on explanatory variable z. The regression estimates and residuals must then satisfy

$$a_i = \lambda_0 + \sum_{z=1}^{k} b_{iz}\lambda_z + \nu_i \qquad (3.28)$$

where by the properties of an orthogonal projection (Ordinary Least Squares regression), the residuals sum to zero, $\sum_{i=1}^{n} \nu_i = 0$, and are orthogonal to the regressors, $\sum_{i=1}^{n} b_{iz}\nu_i = 0$, $z = 1, ..., k$. Thus, we have shown that $(*)$, (i), and (ii) can be satisfied. The last but most important part of the proof is to show that (iii) must hold in the absence of asymptotic arbitrage.

Thus, let us construct a zero-net-investment arbitrage portfolio with the following investment amounts:

$$W_i = \frac{\nu_i}{\sqrt{\sum_{i=1}^{n} \nu_i^2 n}} \tag{3.29}$$

so that greater amounts are invested in assets having the greatest relative expected return deviation. The total arbitrage portfolio return is given by

$$\begin{aligned} \tilde{R}_p &= \sum_{i=1}^{n} W_i \widetilde{R}_i \qquad (3.30) \\ &= \frac{1}{\sqrt{\sum_{i=1}^{n} \nu_i^2 n}} \left[\sum_{i=1}^{n} \nu_i \widetilde{R}_i \right] = \frac{1}{\sqrt{\sum_{i=1}^{n} \nu_i^2 n}} \left[\sum_{i=1}^{n} \nu_i \quad a_i + \sum_{z=1}^{k} b_{iz} \tilde{f}_z + \widetilde{\varepsilon}_i \right) \right] \end{aligned}$$

Since $\sum_{i=1}^{n} b_{iz}\nu_i = 0$, $z = 1, ..., k$, this equals

$$\tilde{R}_p = \frac{1}{\sqrt{\sum_{i=1}^{n} \nu_i^2 n}} \left[\sum_{i=1}^{n} \nu_i \left(a_i + \widetilde{\varepsilon}_i \right) \right] \tag{3.31}$$

Let us calculate this portfolio's mean and variance. Taking expectations, we obtain

$$E\left[\tilde{R}_p\right] = \frac{1}{\sqrt{\sum_{i=1}^{n} \nu_i^2 n}} \left[\sum_{i=1}^{n} \nu_i a_i \right] \tag{3.32}$$

since $E[\widetilde{\varepsilon}_i] = 0$. Substituting in for $a_i = \lambda_0 + \sum_{z=1}^{k} b_{iz}\lambda_z + \nu_i$, we have

$$E\left[\tilde{R}_p\right] = \frac{1}{\sqrt{\sum_{i=1}^{n}\nu_i^2 n}}\left[\lambda_0\sum_{i=1}^{n}\nu_i + \sum_{z=1}^{k}\quad\lambda_z\sum_{i=1}^{n}\nu_i b_{iz}\right) + \sum_{i=1}^{n}\nu_i^2\right] \tag{3.33}$$

and since $\sum_{i=1}^{n}\nu_i = 0$ and $\sum_{i=1}^{n}\nu_i b_{iz} = 0$, this simplifies to

$$E\left[\tilde{R}_p\right] = \frac{1}{\sqrt{\sum_{i=1}^{n}\nu_i^2 n}}\sum_{i=1}^{n}\nu_i^2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\nu_i^2} \tag{3.34}$$

To calculate the portfolio's variance, start by subtracting (3.32) from (3.31):

$$\tilde{R}_p - E\left[\tilde{R}_p\right] = \frac{1}{\sqrt{\sum_{i=1}^{n}\nu_i^2 n}}\left[\sum_{i=1}^{n}\nu_i\widetilde{\varepsilon}_i\right] \tag{3.35}$$

Then, because $E[\widetilde{\varepsilon}_i\widetilde{\varepsilon}_j] = 0$ for $i \neq j$ and $E[\widetilde{\varepsilon}_i^2] = s_i^2$, the portfolio variance is

$$E\left[\left(\tilde{R}_p - E\left[\tilde{R}_p\right]\right)^2\right] = \frac{\sum_{i=1}^{n}\nu_i^2 s_i^2}{n\sum_{i=1}^{n}\nu_i^2} < \frac{\sum_{i=1}^{n}\nu_i^2 S^2}{n\sum_{i=1}^{n}\nu_i^2} = \frac{S^2}{n} \tag{3.36}$$

Thus, as n becomes large ($n \to \infty$), the variance of the portfolio's return goes to zero, that is, the expected return on the portfolio becomes *certain*. This implies that in the limit, the actual return equals the expected return in (3.34):

$$\lim_{n\to\infty}\tilde{R}_p = E\left[\tilde{R}_p\right] = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\nu_i^2} \tag{3.37}$$

and so if there are no asymptotic arbitrage opportunities, this certain return on the portfolio must equal zero. This is equivalent to requiring

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\nu_i^2 = 0 \tag{3.38}$$

which is condition (*iii*).

We see that APT, given by the relation $a_i = \lambda_0 + \sum_{z=1}^{k} b_{iz}\lambda_z$, can be inter-

preted as a multi-beta generalization of CAPM. However, whereas CAPM says that its single beta should be the sensitivity of an asset's return to that of the market portfolio, APT gives no guidance as to what are the economy's multiple underlying risk factors. An empirical application of APT by Nai-Fu Chen, Richard Roll, and Stephen Ross (Chen, Roll, and Ross 1986) assumed that the risk factors were macroeconomic in nature, as proxied by industrial production, expected and unexpected inflation, the spread between long- and short-maturity interest rates, and the spread between high- and low-credit-quality bonds.

Other researchers have tended to select risk factors based on those that provide the "best fit" to historical asset returns.[15] The well-known Eugene Fama and Kenneth French (Fama and French 1993) model is an example of this. Its risk factors are returns on three different portfolios: a market portfolio of stocks (like CAPM), a portfolio that is long the stocks of small firms and short the stocks of large firms, and a portfolio that is long the stocks having high book-to-market ratios (value stocks) and short the stocks having low book-to-market ratios (growth stocks). The latter two portfolios capture the empirical finding that the stocks of smaller firms and those of value firms tend to have higher expected returns than would be predicted solely by the one-factor CAPM model. The Fama-French model predicts that a given stock's expected return is determined by its three betas for these three portfolios.[16] It has been criticized for lacking a theoretical foundation for its risk factors.[17]

However, there have been some attempts to provide a rationale for the Fama-

[15] Gregory Connor and Robert Korajczk (Connor and Korajczyk 1995) survey empirical tests of the APT.

[16] A popular extension of the Fama-French three-factor model is the four-factor model proposed by Mark Carhart (Carhart 1997). His model adds a proxy for stock momentum.

[17] Moreover, some researchers argue that what the model interprets as risk factors may be evidence of market inefficiency. For example, the low returns on growth stocks relative to value stocks may represent market mispricing due to investor overreaction to high growth firms. Josef Lakonishok, Andrei Shleifer, and Robert Vishny (Lakonishok, Shleifer, and Vishny 1994) find that various measures of risk cannot explain the higher average returns of value stocks relative to growth stocks.

French model's good fit of asset returns. Heaton and Lucas (Heaton and Lucas 2000) provide a rationale for the additional Fama-French risk factors. They note that many stockholders may dislike the risks of small-firm and value stocks, the latter often being stocks of firms in financial distress, and thereby requiring higher average returns. They provide empirical evidence that many stockholders are, themselves, entrepreneurs and owners of small businesses, so that their human capital is already subject to the risks of small firms with relatively high probabilities of failure. Hence, these entrepreneurs wish to avoid further exposure to these types of risks.

We will later develop another multibeta asset pricing model, namely Robert Merton's Intertemporal Capital Asset Pricing Model (ICAPM) (Merton 1973a), which is derived from an intertemporal consumer-investor optimization problem. It is a truly dynamic model that allows for changes in state variables that could influence investment opportunities. While the ICAPM is sometimes used to justify the APT, the static (single-period) APT framework may not be compatible with some of the predictions of the more dynamic (multiperiod) ICAPM. In general, the ICAPM allows for changing risk-free rates and predicts that assets' expected returns should be a function of such changing investment opportunities. The model also predicts that an asset's multiple betas are unlikely to remain constant through time, which can complicate deriving estimates of betas from historical data.[18]

3.4 Summary

In this chapter we took a first step in understanding the equilibrium determinants of individual assets' prices and returns. The Capital Asset Pricing Model

[18] Ravi Jagannathan and Zhenyu Wang (Jagannathan and Wang 1996) find that the CAPM better explains stock returns when stocks' betas are permitted to change over time and a proxy for the return on human capital is included in the market portfolio.

(CAPM) was shown to be a natural extension of Markowitz's mean-variance portfolio analysis. However, in addition to deriving CAPM from investor mean-variance risk-preferences, we showed that CAPM and its multifactor generalization Arbitrage Pricing Theory (APT), could result from assumptions of a linear model of asset returns and an absence of arbitrage opportunities.

Arbitrage pricing will arise frequently in subsequent chapters, especially in the context of valuing derivative securities. Furthermore, future chapters will build on our single-period CAPM and APT results to show how equilibrium asset pricing is modified when multiple periods and time-varying asset return distributions are considered.

3.5 Exercises

1. Assume that individual investor k chooses between n risky assets in order to maximize the following utility function:

$$\max_{\{\omega_i^k\}} \quad \overline{R}_k - \frac{1}{\theta_k} V_k$$

where the mean and variance of investor k's portfolio are $\overline{R}_k = \sum_{i=1}^{n} \omega_i^k \overline{R}_i$ and $V_k = \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_i^k \omega_j^k \sigma_{ij}$, respectively, and where $\overline{R}_i$ is the expected return on risky asset i, and σ_{ij} is the covariance between the returns on risky asset i and risky asset j. ω_i^k is investor k's portfolio weight invested in risky asset i, so that $\sum_{i=1}^{n} \omega_i^k = 1$. θ_k is a positive constant and equals investor k's *risk tolerance*.

   (a) Write down the Lagrangian for this problem and show the first-order conditions.

   (b) Rewrite the first-order condition to show that the expected return on

asset i is a linear function of the covariance between risky asset i's return and the return on investor k's optimal portfolio.

(c) Assume that investor k has initial wealth equal to W_k and that there are $k = 1, \ldots, M$ total investors, each with different initial wealth and risk tolerance. Show that the equilibrium expected return on asset i is of a similar form to the first-order condition found in part (b), but depends on the *wealth-weighted risk tolerances* of investors and the *covariance of the return on asset i with the market portfolio.* Hint: begin by multiplying the first order condition in (b) by investor k's wealth times risk tolerance, and then aggregate over all investors.

2. Let the U.S. dollar ($) / Swiss franc (SF) spot exchange rate be $0.68 per SF and the one-year forward exchange rate be $0.70 per SF. The one-year interest rate for borrowing or lending dollars is 6.00 percent.

   (a) What must be the one-year interest rate for borrowing or lending Swiss francs in order for there to be no arbitrage opportunity?

   (b) If the one-year interest rate for borrowing or lending Swiss francs was less than your answer in part (a), describe the arbitrage opportunity.

3. Suppose that the Arbitrage Pricing Theory holds with $k = 2$ risk factors, so that asset returns are given by

$$\widetilde{R}_i = a_i + b_{i1}\tilde{f}_1 + b_{i2}\tilde{f}_2 + \widetilde{\varepsilon}_i$$

where $a_i \cong \lambda_{f0} + b_{i1}\lambda_{f1} + b_{i2}\lambda_{f2}$. Maintain all of the assumptions made in the notes and, in addition, assume that both λ_{f1} and λ_{f2} are positive. Thus, the positive risk premia imply that both of the two orthogonal risk factors are "priced" sources of risk. Now define two new risk factors from the original risk

factors:

$$\widetilde{g}_1 = c_1\widetilde{f}_1 + c_2\widetilde{f}_2$$

$$\widetilde{g}_2 = c_3\widetilde{f}_1 + c_4\widetilde{f}_2$$

Show that there exists a c_1, c_2, c_3, and c_4 such that $\widetilde{g}_1$is orthogonal to $\widetilde{g}_2$, they each have unit variance, and $\lambda_{g1} > 0$, but that $\lambda_{g2} = 0$, where $\lambda_{g1}$and λ_{g2} are the risk premia associated with $\widetilde{g}_1$ and $\widetilde{g}_2$, respectively. In other words, show that any economy with two priced sources of risk can also be described by an economy with one priced source of risk.

Chapter 4

Consumption-Savings Decisions and State Pricing

Previous chapters studied the portfolio choice problem of an individual who maximizes the expected utility of his end-of-period wealth. This specification of an individual's decision-making problem may be less than satisfactory since, traditionally, economists have presumed that individuals derive utility from consuming goods and services, not by possessing wealth per se. Taking this view, our prior analysis can be interpreted as implicitly assuming that the individual consumes only at the end of the single investment period, and all end-of-period wealth is consumed. Utility from the individual consuming some of her initial beginning-of-period wealth was not modeled, so that all initial wealth was assumed to be saved and invested in a portfolio of assets.

In this chapter we consider the more general problem where an individual obtains utility from consuming at both the initial and terminal dates of her decision period and where nontraded labor income also may be received. This allows us to model the individual's initial consumption-savings decision as well

as her portfolio choice decision. In doing so, we can derive relationships between asset prices and the individual's optimal levels of consumption that extend many of our previous results. We introduce the concept of a *stochastic discount factor* that can be used to value the returns on any asset. This stochastic discount factor equals each individual's marginal rate of substitution between initial and end-of-period consumption for each state of nature, that is, each random outcome.

After deriving this stochastic discount factor, we demonstrate that its volatility restricts the feasible excess expected returns and volatilities of all assets. Importantly, we discuss empirical evidence that appears inconsistent with this restriction for standard, time-separable utility functions, casting doubt on the usefulness of a utility-of-consumption-based stochastic discount factor. Fortunately, however, a stochastic discount factor for pricing assets need not rely on this consumption-based foundation. We provide an alternative derivation of a stochastic discount factor based on the assumptions of an absence of arbitrage and *market completeness*. Markets are said to be complete when there are a sufficient number of nonredundant assets whose returns span all states of nature.

The chapter concludes by showing how the stochastic discount factor approach can be modified to derive an asset valuation relationship based on *risk-neutral* probabilities. These probabilities transform the true probabilities of each state of nature to incorporate adjustments for risk premia. Valuation based on risk-neutral probabilities is used extensively to price assets, and this technique will be employed frequently in future chapters.

4.1 Consumption and Portfolio Choices

In this section we introduce an initial consumption-savings decision into an investor's portfolio choice problem. This is done by permitting the individual to derive utility from consuming at the beginning, as well as at the end, of the investment period. The assumptions of our model are as follows.

Let W_0 and C_0 be the individual's initial date 0 wealth and consumption, respectively. At date 1, the end of the period, the individual is assumed to consume all of his wealth which, we denote as C_1. The individual's utility function is defined over beginning- and end-of-period consumption and takes the following form:

$$U(C_0) + \delta E\left[U\left(\widetilde{C}_1\right)\right] \tag{4.1}$$

where δ is a subjective discount factor that reflects the individual's rate of time preference and $E[\cdot]$ is the expectations operator conditional on information at date 0.[1] The multidate specification of utility in expression (4.1) is an example of a *time-separable* utility function. Time separability means that utility at a particular date (say 0 or 1) depends only on consumption at that same date. Later chapters will analyze the implications of time separability and consider generalized multiperiod utility functions that permit utility to depend on past or expected future consumption.

Suppose that the individual can choose to invest in n different assets. Let P_i be the date 0 price per share of asset i, $i = 1, ..., n$, and let X_i be the date 1 random payoff of asset i. For example, a dividend-paying stock might have a

[1] δ is sometimes written as $\frac{1}{1+\rho}$ where ρ is the rate of time preference. A value of $\delta < 1$ ($\rho > 0$) reflects impatience on the part of the individual, that is, a preference for consuming early. A more general two-date utility function could be expressed as $U_0(C_o) + E[U_1(C_1)]$ where U_0 and U_1 are any different increasing, concave functions of consumption. Our presentation assumes $U_1(C) = \delta U_0(C)$, but the qualitative results we derive also hold for the more general specification.

date 1 random payoff of $\widetilde{X}_i = \widetilde{P}_{1i} + \widetilde{D}_{1i}$, where $\widetilde{P}_{1i}$ is the date 1 stock price and $\widetilde{D}_{1i}$ is the stock's dividend paid at date 1. Alternatively, for a coupon-paying bond, $\widetilde{P}_{1i}$ would be the date 1 bond price and $\widetilde{D}_{1i}$ would be the bond's coupon paid at date 1.[2] Given this definition, we can also define $R_i \equiv X_i/P_i$ to be the random return on asset i. The individual may also receive labor income of y_0 at date 0 and random labor income of y_1 at date 1.[3] If ω_i is the proportion of date 0 savings that the individual chooses to invest in asset i, then his intertemporal budget constraint is

$$C_1 = y_1 + (W_0 + y_0 - C_0) \sum_{i=1}^{n} \omega_i R_i \tag{4.2}$$

where $(W_0 + y_0 - C_0)$ is the individual's date 0 savings. The individual's maximization problem can then be stated as

$$\max_{C_0, \{\omega_i\}} U(C_0) + \delta E[U(C_1)] \tag{4.3}$$

subject to equation (4.2) and the constraint $\sum_{i=1}^{n} \omega_i = 1$. The first-order conditions with respect to C_0 and the ω_i, $i = 1, ..., n$ are

$$U'(C_0) - \delta E\left[U'(C_1) \sum_{i=1}^{n} \omega_i R_i\right] = 0 \tag{4.4}$$

$$\delta E[U'(C_1) R_i] - \lambda = 0, \quad i = 1, ..., n \tag{4.5}$$

where $\lambda \equiv \lambda'/(W_0 + y_0 - C_0)$ and λ' is the Lagrange multiplier for the constraint $\sum_{i=1}^{n} \omega_i = 1$. The first-order conditions in (4.5) describe how the in-

[2] The coupon payment would be uncertain if default on the payment is possible and/or the coupon is not fixed but floating (tied to a market interest rate).

[3] There is an essential difference between tangible wealth, W, and wage income, y. The present value of wage income, which is referred to as "human capital," is assumed to be a nontradeable asset. The individual can rebalance his tangible wealth to change his holdings of marketable assets, but his endowment of human capital (and its cashflows in the form of wage income) is assumed to be fixed.

vestor chooses between different assets. Substitute out for λ and one obtains

$$E\left[U'\left(C_1\right)R_i\right] = E\left[U'\left(C_1\right)R_j\right] \tag{4.6}$$

for any two assets, i and j. Equation (4.6) tells us that the investor trades off investing in asset i for asset j until their expected marginal utility-weighted returns are equal. If this were not the case, the individual could raise his total expected utility by investing more in assets whose marginal utility-weighted returns were relatively high and investing less in assets whose marginal utility-weighted returns were low.

How does the investor act to make the optimal equality of expected marginal utility-weighted returns in (4.6) come about? Note from (4.2) that C_1 becomes more positively correlated with R_i the greater is ω_i. Thus, the greater asset i's portfolio weight, the lower will be $U'\left(C_1\right)$ when R_i is high due to the concavity of utility. Hence, as ω_i becomes large, smaller marginal utility weights multiply the high realizations of asset i's return, and $E\left[U'\left(C_1\right)R_i\right]$ falls. Intuitively, this occurs because the investor becomes more undiversified by holding a larger proportion of asset i. By adjusting the portfolio weights for asset i and each of the other $n-1$ assets, the investor changes the random distribution of C_1 in a way that equalizes $E\left[U'\left(C_1\right)R_k\right]$ for all assets $k = 1, ..., n$, thereby attaining the desired level of diversification.

Another result of the first-order conditions involves the intertemporal allocation of resources. Substituting (4.5) into (4.4) gives

$$\begin{aligned} U'\left(C_0\right) &= \delta E\left[U'\left(C_1\right)\sum_{i=1}^{n}\omega_i R_i\right] = \sum_{i=1}^{n}\omega_i \delta E\left[U'\left(C_1\right)R_i\right] \\ &= \sum_{i=1}^{n}\omega_i \lambda = \lambda \end{aligned} \tag{4.7}$$

Therefore, substituting $\lambda = U'(C_0)$, the first-order conditions in (4.5) can be written as

$$\delta E\left[U'(C_1) R_i\right] = U'(C_0), \quad i = 1, ..., n \tag{4.8}$$

or, since $R_i = X_i/P_i$,

$$P_i U'(C_0) = \delta E\left[U'(C_1) X_i\right], \quad i = 1, ..., n \tag{4.9}$$

Equation (4.9) has an intuitive meaning and, as will be shown in subsequent chapters, generalizes to multiperiod consumption and portfolio choice problems. It says that when the investor is acting optimally, he invests in asset i until the loss in marginal utility of giving up P_i dollars at date 0 just equals the expected marginal utility of receiving the random payoff of X_i at date 1. To see this more clearly, suppose that one of the assets pays a risk-free return over the period. Call it asset f so that R_f is the risk-free return (1 plus the risk-free interest rate). For the risk-free asset, equation (4.9) can be rewritten as

$$U'(C_0) = R_f \delta E\left[U'(C_1)\right] \tag{4.10}$$

which states that the investor trades off date 0 for date 1 consumption until the marginal utility of giving up \$1 of date 0 consumption just equals the expected marginal utility of receiving $\$R_f$ of date 1 consumption. For example, suppose that utility is of a constant relative-risk-aversion form: $U(C) = C^\gamma/\gamma$, for $\gamma < 1$. Then equation (4.10) can be rewritten as

$$\frac{1}{R_f} = \delta E\left[\left(\frac{C_0}{C_1}\right)^{1-\gamma}\right] \tag{4.11}$$

Hence, when the interest rate is high, so will be the expected growth in consump-

tion. For the special case of there being only one risk-free asset and nonrandom labor income, so that C_1 is nonstochastic, equation (4.11) becomes

$$R_f = \frac{1}{\delta}\left(\frac{C_1}{C_0}\right)^{1-\gamma} \tag{4.12}$$

Taking logs of both sides of the equation, we obtain

$$\ln(R_f) = -\ln\delta + (1-\gamma)\ln\left(\frac{C_1}{C_0}\right) \tag{4.13}$$

Since $\ln(R_f)$ is the continuously compounded, risk-free interest rate and $\ln(C_1/C_0)$ is the growth rate of consumption, then we can define the elasticity of intertemporal substitution, ϵ, as

$$\epsilon \equiv \frac{\partial \ln(C_1/C_0)}{\partial \ln(R_f)} = \frac{1}{1-\gamma} \tag{4.14}$$

Hence, with power (constant relative-risk-aversion) utility, ϵ is the reciprocal of the coefficient of relative risk aversion. That is, the single parameter γ determines both risk aversion and the rate of intertemporal substitution.[4] When $0 < \gamma < 1$, ϵ exceeds unity and a higher interest rate raises second-period consumption more than one-for-one. This implies that if utility displays less risk aversion than logarithmic utility, this individual increases his savings as the interest rate rises. Conversely, when $\gamma < 0$, then $\epsilon < 1$ and a rise in the interest rate raises second-period consumption less than one-for-one, implying that such an individual decreases her initial savings when the return to savings is higher. For the logarithmic utility individual ($\gamma = 0$ and therefore, $\epsilon = 1$), a change in the interest rate has no effect on savings. These results can be interpreted as

[4] An end-of-chapter exercise shows that this result extends to an environment with risky assets. In Chapter 14, we will examine a recursive utility generalization of multiperiod power utility for which the elasticity of intertemporal substitution is permitted to differ from the inverse of the coefficient of relative risk aversion. There these two characteristics of multiperiod utility are modeled by separate parameters.

an individual's response to two effects from an increase in interest rates. The first is a *substitution effect* that raises the return from transforming current consumption into future consumption. This higher benefit from initial savings provides an incentive to do more of it. The second effect is an *income effect* due to the greater return that is earned on a given amount of savings. This makes the individual better off and, ceteris paribus, would raise consumption in both periods. Hence, initial savings could fall and still lead to greater consumption in the second period. For $\epsilon > 1$, the substitution effect outweighs the income effect, while the reverse occurs when $\epsilon < 1$. When $\epsilon = 1$, the income and substitution effects exactly offset each other.

A main insight of this section is that an individual's optimal portfolio of assets is one where the assets' expected marginal utility-weighted returns are equalized. If this were not the case, the individual's expected utility could be raised by investing more (*less*) in assets whose average marginal utility-weighted returns are relatively high (*low*). It was also demonstrated that an individual's optimal consumption-savings decision involves trading off higher current marginal utility of consuming for higher expected future marginal utility obtainable from invested saving.

4.2 An Asset Pricing Interpretation

Until now, we have analyzed the consumption-portfolio choice problem of an individual investor. For such an exercise, it makes sense to think of the individual taking the current prices of all assets and the distribution of their payoffs as given when deciding on his optimal consumption-portfolio choice plan. Importantly, however, the first-order conditions we have derived might be re-interpreted as asset pricing relationships. They can provide insights regarding the connection between individuals' consumption behavior and the distribution of asset

returns.

To see this, let us begin by rewriting equation (4.9) as

$$\begin{aligned} P_i &= E\left[\frac{\delta U'(C_1)}{U'(C_0)} X_i\right] \\ &= E\left[m_{01} X_i\right] \end{aligned} \tag{4.15}$$

where $m_{01} \equiv \delta U'(C_1)/U'(C_0)$ is the marginal rate of substitution between initial and end-of-period consumption. For any individual who can trade freely in asset i, equation (4.15) provides a condition that equilibrium asset prices must satisfy. Condition (4.15) appears in the form of an asset pricing formula. The current asset price, P_i, is an expected discounted value of its payoffs, where the discount factor, m_{01}, is a random quantity because it depends on the random level of future consumption. Hence, m_{01} is also referred to as the *stochastic discount factor* for valuing asset returns. In states of nature where future consumption turns out to be high (due to high asset portfolio returns or high labor income), marginal utility, $U'(C_1)$, is low and the asset's payoffs in these states are not highly valued. Conversely, in states where future consumption is low, marginal utility is high so that the asset's payoffs in these states are much desired. This insight explains why m_{01} is also known as the *state price deflator*. It provides a different discount factor (deflator) for different states of nature.

It should be emphasized that the stochastic discount factor, m_{01}, is the same for all assets that a particular investor can hold. It prices these assets' payoffs only by differentiating in which state of nature the payoff is made. Since m_{01} provides the core, or kernel, for pricing all risky assets, it is also referred to as the *pricing kernel*. Note that the random realization of m_{01} may differ across investors because of differences in random labor income that can cause the random distribution of C_1 to vary across investors. Nonetheless, the expected

product of the pricing kernel and asset i's payoff, $E[m_{01}X_i]$, will be the same for all investors who can trade in asset i.

4.2.1 Real versus Nominal Returns

In writing down the individual's consumption-portfolio choice problem, we implicitly assumed that returns are expressed in real, or *purchasing power*, terms; that is, returns should be measured after adjustment for inflation. The reason is that an individual's utility should depend on the real, not nominal (currency denominated), value of consumption. Therefore, in the budget constraint (4.2), if C_1 denotes real consumption, then asset returns and prices (as well as labor income) need to be real values. Thus, if P_i^N and X_i^N are the initial price and end-of-period payoff measured in currency units (nominal terms), we need to deflate them by a price index to convert them to real quantities. Letting CPI_t denote the consumer price index at date t, the pricing relationship in (4.15) becomes

$$\frac{P_i^N}{CPI_0} = E\left[\frac{\delta U'(C_1)}{U'(C_0)}\frac{X_i^N}{CPI_1}\right] \tag{4.16}$$

or if we define $I_{ts} = CPI_s/CPI_t$ as 1 plus the inflation rate between dates t and s, equation (4.16) can be rewritten as

$$\begin{aligned} P_i^N &= E\left[\frac{1}{I_{01}}\frac{\delta U'(C_1)}{U'(C_0)}X_i^N\right] \\ &= E\left[M_{01}X_i^N\right] \end{aligned} \tag{4.17}$$

where $M_{01} \equiv (\delta/I_{01})\,U'(C_1)/U'(C_0)$ is the stochastic discount factor (pricing kernel) for discounting nominal returns. Hence, this nominal pricing kernel is simply the real pricing kernel, m_{01}, discounted at the (random) rate of inflation

between dates 0 and 1.

4.2.2 Risk Premia and the Marginal Utility of Consumption

The relation in equation (4.15) can be rewritten to shed light on an asset's risk premium. Dividing each side of (4.15) by P_i results in

$$\begin{aligned} 1 &= E\left[m_{01}R_i\right] \\ &= E\left[m_{01}\right]E\left[R_i\right] + Cov\left[m_{01}, R_i\right] \\ &= E\left[m_{01}\right]\left(E\left[R_i\right] + \frac{Cov\left[m_{01}, R_i\right]}{E\left[m_{01}\right]}\right) \end{aligned} \tag{4.18}$$

Recall from (4.10) that for the case of a risk-free asset, $E\left[\delta U'\left(C_1\right)/U'\left(C_0\right)\right] = E\left[m_{01}\right] = 1/R_f$. Then (4.18) can be rewritten as

$$R_f = E\left[R_i\right] + \frac{Cov\left[m_{01}, R_i\right]}{E\left[m_{01}\right]} \tag{4.19}$$

or

$$\begin{aligned} E\left[R_i\right] &= R_f - \frac{Cov\left[m_{01}, R_i\right]}{E\left[m_{01}\right]} \\ &= R_f - \frac{Cov\left[U'\left(C_1\right), R_i\right]}{E\left[U'\left(C_1\right)\right]} \end{aligned} \tag{4.20}$$

Equation (4.20) states that the risk premium for asset i equals the negative of the covariance between the marginal utility of end-of-period consumption and the asset return divided by the expected end-of-period marginal utility of consumption. If an asset pays a higher return when consumption is high, its return has a negative covariance with the marginal utility of consumption, and

therefore the investor demands a positive risk premium over the risk-free rate.

Conversely, if an asset pays a higher return when consumption is low, so that its return positively covaries with the marginal utility of consumption, then it has an expected return less than the risk-free rate. Investors will be satisfied with this lower return because the asset is providing a hedge against low consumption states of the world; that is, it is helping to smooth consumption across states.

4.2.3 The Relationship to CAPM

Now suppose there exists a portfolio with a random return of $\widetilde{R}_m$ that is perfectly negatively correlated with the marginal utility of date 1 consumption, $U'\left(\widetilde{C}_1\right)$, implying that it is also perfectly negatively correlated with the pricing kernel, m_{01}:

$$U'(\tilde{C}_1) \,=\, \kappa_0 - \kappa\, \widetilde{R}_m, \quad \kappa_0 > 0,\ \kappa > 0 \tag{4.21}$$

Then this implies

$$Cov[U'(C_1),\, R_m] \,=\, -\kappa\, Cov[R_m, R_m] \,=\, -\,\kappa\, Var[R_m] \tag{4.22}$$

and

$$Cov[U'(C_1),\, R_i] \,=\, -\kappa\, Cov[R_m,\, R_i] \tag{4.23}$$

For the portfolio having return $\widetilde{R}_m$, the risk premium relation (4.20) is

$$E[R_m] \,=\, R_f - \frac{Cov[U'(C_1),\, R_m]}{E[U'(C_1)]} \,=\, R_f + \frac{\kappa\, Var[R_m]}{E[U'(C_1)]} \tag{4.24}$$

Using (4.20) and (4.24) to substitute for $E[U'(C_1)]$, and using (4.23), we obtain

$$\frac{E[R_m] - R_f}{E[R_i] - R_f} \,=\, \frac{\kappa\, Var[R_m]}{\kappa Cov[R_m,\, R_i]} \tag{4.25}$$

and rearranging:

$$E[R_i] - R_f \;=\; \frac{Cov[R_m, R_i]}{Var[R_m]}\,(E[R_m] - R_f) \tag{4.26}$$

or

$$E[R_i] \;=\; R_f + \beta_i\,(E[R_m] - R_f) \tag{4.27}$$

So we obtain the CAPM if the return on the market portfolio is perfectly negatively correlated with the marginal utility of end-of-period consumption, that is, perfectly negatively correlated with the pricing kernel. Note that for an arbitrary distribution of asset returns and nonrandom labor income, this will always be the case if utility is quadratic, because marginal utility is linear in consumption and consumption also depends linearly on the market's return. In addition, for the case of general utility, normally distributed asset returns, and nonrandom labor income, marginal utility of end-of-period consumption is also perfectly negatively correlated with the return on the market portfolio, because each investor's optimal portfolio is simply a combination of the market portfolio and the (nonrandom) risk-free asset. Thus, consistent with Chapters 2 and 3, under the assumptions needed for mean-variance analysis to be equivalent with expected utility maximization, asset returns satisfy the CAPM.

4.2.4 Bounds on Risk Premia

Another implication of the stochastic discount factor is that it places bounds on the means and standard deviations of individual securities and, therefore, determines an efficient frontier. To show this, rewrite the first line in equation (4.20) as

$$E\left[R_i\right] = R_f - \rho_{m_{01},R_i}\frac{\sigma_{m_{01}}\sigma_{R_i}}{E\left[m_{01}\right]} \tag{4.28}$$

where $\sigma_{m_{01}}$, σ_{R_i}, and ρ_{m_{01},R_i} are the standard deviation of the discount factor, the standard deviation of the return on asset i, and the correlation between the discount factor and the return on asset i, respectively. Rearranging (4.28) leads to

$$\frac{E[R_i] - R_f}{\sigma_{R_i}} = -\rho_{m_{01},R_i} \frac{\sigma_{m_{01}}}{E[m_{01}]} \tag{4.29}$$

The left-hand side of (4.29) is the Sharpe ratio for asset i. Since $-1 \leq \rho_{m_{01},R_i} \leq 1$, we know that

$$\left| \frac{E[R_i] - R_f}{\sigma_{R_i}} \right| \leq \frac{\sigma_{m_{01}}}{E[m_{01}]} = \sigma_{m_{01}} R_f \tag{4.30}$$

This equation was derived by Robert Shiller (Shiller 1982), was generalized by Lars Hansen and Ravi Jagannathan (Hansen and Jagannathan 1991), and is known as a Hansen-Jagannathan bound. Given an asset's Sharpe ratio and the risk-free rate, equation (4.30) sets a lower bound on the volatility of the economy's stochastic discount factor. Conversely, given the volatility of the discount factor, equation (4.30) sets an upper bound on the maximum Sharpe ratio that any asset, or portfolio of assets, can attain.

If there exists an asset (or portfolio of assets) whose return is perfectly negatively correlated with the discount factor, m_{01}, then the bound in (4.30) holds with equality. As we just showed in equations (4.21) to (4.27), such a situation implies the CAPM, so that the slope of the capital market line, $S_e \equiv \frac{E[R_m]-R_f}{\sigma_{R_m}}$, equals $\sigma_{m_{01}} R_f$. Thus, the slope of the capital market line, which represents (efficient) portfolios that have a maximum Sharpe ratio, can be related to the standard deviation of the discount factor.

The inequality in (4.30) has empirical implications. $\sigma_{m_{01}}$ can be estimated if we could observe an individual's consumption stream and if we knew his or

her utility function. Then, according to (4.30), the Sharpe ratio of any portfolio of traded assets should be less than or equal to $\sigma_{m_{01}}/E[m_{01}]$. For power utility, $U(C) = C^\gamma/\gamma$, $\gamma < 1$, so that $m_{01} \equiv \delta (C_1/C_0)^{\gamma-1} = \delta e^{(\gamma-1)\ln(C_1/C_0)}$. If C_1/C_0 is assumed to be lognormally distributed, with parameters μ_c and σ_c, then

$$\begin{aligned}
\frac{\sigma_{m_{01}}}{E[m_{01}]} &= \frac{\sqrt{Var\left[e^{(\gamma-1)\ln(C_1/C_0)}\right]}}{E\left[e^{(\gamma-1)\ln(C_1/C_0)}\right]} \\
&= \frac{\sqrt{E\left[e^{2(\gamma-1)\ln(C_1/C_0)}\right] - E\left[e^{(\gamma-1)\ln(C_1/C_0)}\right]^2}}{E\left[e^{(\gamma-1)\ln(C_1/C_0)}\right]} \\
&= \sqrt{E\left[e^{2(\gamma-1)\ln(C_1/C_0)}\right]/E\left[e^{(\gamma-1)\ln(C_1/C_0)}\right]^2 - 1} \\
&= \sqrt{e^{2(\gamma-1)\mu_c + 2(\gamma-1)^2\sigma_c^2}/e^{2(\gamma-1)\mu_c + (\gamma-1)^2\sigma_c^2} - 1} = \sqrt{e^{(\gamma-1)^2\sigma_c^2} - 1} \\
&\approx (1-\gamma)\,\sigma_c \qquad (4.31)
\end{aligned}$$

where in the fourth line of (4.31), the expectations are evaluated assuming C_1 is lognormally distributed.[5] Hence, with power utility and lognormally distributed consumption, we have

$$\left|\frac{E[R_i] - R_f}{\sigma_{R_i}}\right| \leq (1-\gamma)\,\sigma_c \qquad (4.32)$$

Suppose, for example, that R_i is the return on a broadly diversified portfolio of U.S. stocks, such as the S&P 500. Over the last 75 years, this portfolio's annual real return in excess of the risk-free (U.S. Treasury bill) interest rate has averaged 8.3 percent, suggesting $E[R_i] - R_f = 0.083$. The portfolio's annual standard deviation has been approximately $\sigma_{R_i} = 0.17$, implying a Sharpe ratio of $\frac{E[R_i]-R_f}{\sigma_{R_i}} = 0.49$. Assuming a "representative agent" and using per

[5] The fifth line of (4.31) is based on taking a two-term approximation of the series $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ...$, which is reasonable when x is a small positive number.

capita U.S. consumption data to estimate the standard deviation of consumption growth, researchers have come up with annualized estimates of σ_c between 0.01 and 0.0386.[6] Thus, even if a diversified portfolio of U.S. stocks was an efficient portfolio of risky assets, so that equation (4.32) held with equality, it would imply a value of $\gamma = 1 - \left(\frac{E[R_i]-R_f}{\sigma_{R_i}}\right)/\sigma_c$ between -11.7 and -48.[7] Since reasonable levels of risk aversion estimated from other sources imply values of γ much smaller in magnitude, say in the range of -1 to -5, the inequality (4.32) appears not to hold for U.S. stock market data and standard specifications of utility.[8] In other words, consumption appears to be too smooth (σ_c is too low) relative to the premium that investors demand for holding stocks. This inconsistency between theory and empirical evidence was identified by Rajnish Mehra and Edward Prescott (Mehra and Prescott 1985) and is referred to as the *equity premium puzzle*. Attempts to explain this puzzle have involved using different specifications of utility and questioning whether the ex-post sample mean of U.S. stock returns is a good estimate of the a priori expected return on U.S. stocks.[9]

Even if one were to accept a high degree of risk aversion in order to fit the historical equity premium, additional problems may arise because this high risk aversion could imply an unreasonable value for the risk-free return, R_f. Under our maintained assumptions and using (4.10), the risk-free return satisfies

[6] See John Y. Campbell (Campbell 1999) and Stephen G. Cecchetti, Pok-Sam Lam, and Nelson C. Mark (Cecchetti, Lam, and Mark 1994).

[7] If the stock portfolio were less than efficient, so that a strict inequality held in (4.32), the magnitude of the risk-aversion coefficient would need to be even higher.

[8] Rajnish Mehra and Edward Prescott (Mehra and Prescott 1985) survey empirical work, finding values of γ of -1 or more (equivalent to coefficients of relative risk aversion, $1-\gamma$, of 2 or less).

[9] Jeremy J. Siegel and Richard H. Thaler (Siegel and Thaler 1997) review this literature. It should be noted that recent survey evidence from academic financial economists (Welch 2000) finds that a consensus believes that the current equity risk premium is significantly lower than the historical average. Moreover, at the begining of 2006, the Federal Reserve Bank of Philadelphia's *Survey of Professional Forecasters* found that the median predicted annual returns over the next decade on the S&P 500 stock portfolio, the 10-year U.S. Treasury bond, and the 3-month U.S. Treasury bill are 7.00%, 5.00%, and 4.25%, respectively. This implies a much lower equity risk premium (7.00% - 4.25% = 2.75%) compared to the historical average difference between stocks and bills of 8.3%.

$$\begin{aligned} \frac{1}{R_f} &= E\left[m_{01}\right] \\ &= \delta E\left[e^{(\gamma-1)\ln(C_1/C_0)}\right] \\ &= \delta e^{(\gamma-1)\mu_c+\frac{1}{2}(\gamma-1)^2\sigma_c^2} \end{aligned} \tag{4.33}$$

and therefore

$$\ln\left(R_f\right) = -\ln\left(\delta\right) + \left(1-\gamma\right)\mu_c - \frac{1}{2}\left(1-\gamma\right)^2\sigma_c^2 \tag{4.34}$$

If we set $\delta = 0.99$, reflecting a 1 percent rate of time preference, and $\mu_c = 0.018$, which is the historical average real growth of U.S. per capita consumption, then a value of $\gamma = -11$ and $\sigma_c = 0.036$ implies

$$\begin{aligned} \ln\left(R_f\right) &= -\ln\left(\delta\right) + \left(1-\gamma\right)\mu_c - \frac{1}{2}\left(1-\gamma\right)^2\sigma_c^2 \\ &= 0.01 + 0.216 - 0.093 = 0.133 \end{aligned} \tag{4.35}$$

which is a real risk-free interest rate of 13.3 percent. Since short-term real interest rates have averaged about 1 percent in the United States, we end up with a *risk-free rate puzzle.*

The notion that assets can be priced using a stochastic discount factor, m_{01}, is attractive because the discount factor is independent of the asset being priced: it can be used to price any asset no matter what its risk. We derived this discount factor from a consumption-portfolio choice problem and, in this context, showed that it equaled the marginal rate of substitution between current and end-of-period consumption. However, the usefulness of this approach is in doubt since empirical evidence using aggregate consumption data and standard spec-

ifications of utility appears inconsistent with the discount factor equaling the marginal rate of substitution.[10] Fortunately, a general pricing relationship of the form $P_i = E_0[m_{01}X_i]$ can be shown to hold without assuming that m_{01} represents a marginal rate of substitution. Rather, it can be derived using alternative assumptions. This is the subject of the next section.

4.3 Market Completeness, Arbitrage, and State Pricing

We need not assume a consumption-portfolio choice structure to derive a stochastic discount factor pricing formula. Instead, our derivation can be based on the assumptions of a complete market and the absence of arbitrage, an approach pioneered by Kenneth Arrow and Gerard Debreu.[11] With these alternative assumptions, one can show that a law of one price holds and that a unique stochastic discount factor exists. This new approach makes transparent the derivation of relative pricing relationships and is an important technique for valuing contingent claims (derivatives).

4.3.1 Complete Markets Assumptions

To illustrate, suppose once again that an individual can freely trade in n different assets. Also, let us assume that there are a finite number of end-of-period states of nature, with state s having probability π_s.[12] Let X_{si} be the cashflow generated by one share (unit) of asset i in state s. Also assume that there are k states of nature and n assets. The following vector describes the payoffs to

[10] As will be shown in Chapter 14, some specifications of time-inseparable utility can improve the consumption-based stochastic discount factor's ability to explain asset prices.

[11] See Kenneth Arrow (Arrow 1953) reprinted in (Arrow 1964) and Gerard Debreu (Debreu 1959).

[12] As is discussed later, this analysis can be extended to the case of an infinite number of states.

financial asset i:

$$X_i = \begin{bmatrix} X_{1i} \\ \vdots \\ X_{ki} \end{bmatrix} \tag{4.36}$$

Thus, the per-share cashflows of the universe of all assets can be represented by the $k \times n$ matrix

$$X = \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{k1} & \cdots & X_{kn} \end{bmatrix} \tag{4.37}$$

We will assume that $n = k$ and that X is of full rank. This implies that the n assets *span* the k states of nature, an assumption that indicates a complete market. We would still have a complete market (and, as we will show, unique state-contingent prices) if $n > k$, as long as the payoff matrix X has rank k. If the number of assets exceeds the number of states, some assets are redundant; that is, their cashflows in the k states are linear combinations of others. In such a situation, we could reduce the number of assets to k by combining them into k linearly independent (portfolios of) assets.

An implication of the assumption that the assets' returns span the k states of nature is that an individual can purchase amounts of the k assets so that she can obtain target levels of end-of-period wealth in each of the states. To show this complete markets result, let W denote an arbitrary $k \times 1$ vector of end-of-period levels of wealth:

$$W = \begin{bmatrix} W_1 \\ \vdots \\ W_k \end{bmatrix} \tag{4.38}$$

where W_s is the level of wealth in state s. To obtain W, at the initial date

the individual needs to purchase shares in the k assets. Let the vector $N = [N_1 \ldots N_k]'$ be the number of shares purchased of each of the k assets. Hence, N must satisfy

$$XN = W \tag{4.39}$$

Because X is a nonsingular matrix of rank k, its inverse exists so that

$$N = X^{-1}W \tag{4.40}$$

Hence, because the assets' payoffs span the k states, arbitrary levels of wealth in the k states can be attained if initial wealth is sufficient to purchase the required shares, N. Denoting $P = [P_1 \ldots P_k]'$ as the $k \times 1$ vector of beginning-of-period, per-share prices of the k assets, then the amount of initial wealth required to produce the target level of wealth given in (4.38) is simply $P'N$.

4.3.2 Arbitrage and State Prices

Given our assumption of complete markets, the absence of arbitrage opportunities implies that the price of a new, redundant security or contingent claim can be valued based on the prices of the original k securities. For example, suppose a new asset pays a vector of end-of-period cashflows of W. In the absence of arbitrage, its price must be $P'N$. If its price exceeded $P'N$, an arbitrage would be to sell this new asset and purchase the original k securities in amounts N. Since the end-of-period liability from selling the security is exactly offset by the returns received from the k original securities, the arbitrage profit equals the difference between the new asset's price and $P'N$. Conversely, if the new asset's price were less than $P'N$, an arbitrage would be to purchase the new asset and sell the portfolio N of the k original securities.

Let's apply this concept of complete markets, no-arbitrage pricing to the

special case of a security that has a payoff of 1 in state s and 0 in all other states. Such a security is referred to as a *primitive*, *elementary*, or *Arrow-Debreu* security. Specifically, elementary security "s" has the vector of cashflows

$$e_s = \begin{bmatrix} W_1 \\ \vdots \\ W_{s-1} \\ W_s \\ W_{s+1} \\ \vdots \\ W_k \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4.41}$$

Let p_s be the beginning-of-period price of elementary security s, that is, the price of receiving 1 in state s. Then as we just showed, its price in terms of the payoffs and prices of the original k assets must equal

$$p_s = P'X^{-1}e_s, \quad s = 1, ..., k \tag{4.42}$$

so that a unique set of state prices exists in a complete market.[13] Furthermore, we would expect that these elementary state prices should each be positive, since a unit amount of wealth received in any state will have a value greater than zero whenever individuals are assumed to be nonsatiated.[14] Hence the equations in (4.42) along with the conditions $p_s > 0 \ \forall s$ restrict the payoffs, X, and the prices, P, of the original k securities.

We can now derive a stochastic discount factor formula by considering the

[13] If markets were incomplete, for example, if n were the rank of X and $k > n$, then state prices would not be uniquely determined by the absence of arbitrage. The no-arbitrage conditions would place only n linear restrictions on the set of k prices, implying that there could be an infinity of possible state prices.

[14] This would be the case whenever individuals' marginal utilities are positive for all levels of end-of-period consumption.

value of any other security or contingent claim in terms of these elementary state security prices. Note that the portfolio composed of the sum of all elementary securities gives a cashflow of 1 unit with certainty. The price of this portfolio defines the risk-free return, R_f, by the relation

$$\sum_{s=1}^{k} p_s = \frac{1}{R_f} \tag{4.43}$$

In general, let there be some multicashflow asset, a, whose cashflow paid in state s is X_{sa}. In the absence of arbitrage, its price, P_a, must equal

$$P_a = \sum_{s=1}^{k} p_s \, X_{sa} \tag{4.44}$$

Note that the relative pricing relationships that we have derived did not require using information on the state probabilities. However, let us now introduce these probabilities to see their relationship to state prices and the stochastic discount factor. Define $m_s \equiv p_s/\pi_s$ to be the price of elementary security s divided by the probability that state s occurs. Note that if, as was argued earlier, a sensible equilibrium requires $p_s > 0 \; \forall s$, then $m_s > 0 \; \forall s$ when there is a positive probability of each state occurring. Using this new definition, equation (4.44) can be written as

$$\begin{aligned} P_a &= \sum_{s=1}^{k} \pi_s \frac{p_s}{\pi_s} \, X_{sa} \\ &= \sum_{s=1}^{k} \pi_s m_s \, X_{sa} \\ &= E\left[m \, X_a\right] \end{aligned} \tag{4.45}$$

where m denotes a stochastic discount factor whose expected value is $\sum_{s=1}^{k} \pi_s m_s$,

and X_a is the random cashflow of the multicashflow asset a. Equation (4.45) shows that the stochastic discount factor equals the prices of the elementary securities normalized by their state probabilities. Hence, we have shown that in a complete market that lacks arbitrage opportunities, a unique, positive-valued stochastic discount factor exists. When markets are incomplete, the absence of arbitrage, alone, cannot determine the stochastic discount factor. One would need to impose additional conditions, such as the previous section's assumptions on the form of individuals' utility, in order to determine the stochastic discount factor. For example, if different states of nature led to different realizations of an individual's nontraded labor income, and there did not exist assets that could span or insure against this wage income, then a unique stochastic discount factor may not exist. In this case of market incompleteness, a utility-based derivation of the stochastic discount factor may be required for asset pricing.

While the stochastic discount factor relationship of equation (4.45) is based on state prices derived from assumptions of market completeness and the absence of arbitrage, it is interesting to interpret these state prices in terms of the previously derived consumption-based discount factor. Note that since $p_s = \pi_s m_s$, the price of the elementary security paying 1 in state s is higher the greater the likelihood of the state s occurring and the greater the stochastic discount factor for state s. In terms of the consumption-based model, $m_s = \delta U'(C_{1s}) / U'(C_0)$ where C_{1s} is the level of consumption at date 1 in state s. Hence, the state s price, p_s, is greater when C_{1s} is low; that is, state s is a low consumption state, such as an economic recession.

4.3.3 Risk-Neutral Probabilities

The state pricing relationship of equation (4.44) can be used to develop an important alternative formula for pricing assets. Define $\widehat{\pi}_s \equiv p_s R_f$ as the price

of elementary security s times the risk-free return. Then

$$\begin{aligned} P_a &= \sum_{s=1}^{k} p_s \, X_{sa} \\ &= \frac{1}{R_f} \sum_{s=1}^{k} p_s R_f \, X_{sa} \\ &= \frac{1}{R_f} \sum_{s=1}^{k} \widehat{\pi}_s \, X_{sa} \end{aligned} \tag{4.46}$$

Now these $\widehat{\pi}_s$, $s = 1, ..., k$, have the characteristics of probabilities because they are positive, $\widehat{\pi}_s = p_s / \sum_{s=1}^{k} p_s > 0$, and they sum to 1, $\sum_{s=1}^{k} \widehat{\pi}_s = R_f \sum_{s=1}^{k} p_s = R_f / R_f = 1$. Using this insight, we can rewrite equation (4.46) as

$$\begin{aligned} P_a &= \frac{1}{R_f} \sum_{s=1}^{k} \widehat{\pi}_s \, X_{sa} \\ &= \frac{1}{R_f} \widehat{E}\left[X_a\right] \end{aligned} \tag{4.47}$$

where $\widehat{E}\left[\cdot\right]$ denotes the expectation operator evaluated using the "pseudo" probabilities $\widehat{\pi}_s$ rather than the true probabilities π_s. Since the expectation in (4.47) is discounted by the risk-free return, we can recognize $\widehat{E}\left[X_a\right]$ as the certainty equivalent expectation of the cashflow X_a. In comparison to the stochastic discount factor approach, the formula works by modifying the probabilities of the cashflows in each of the different states, rather than discounting the cashflows by a different discount factor. To see this, note that since $m_s \equiv p_s / \pi_s$ and $R_f = 1 / E\left[m\right]$, $\widehat{\pi}_s$ can be written as

$$\begin{aligned} \widehat{\pi}_s &= R_f m_s \pi_s \\ &= \frac{m_s}{E[m]} \pi_s \end{aligned} \tag{4.48}$$

so that the pseudo probability transforms the true probability by multiplying by the ratio of the stochastic discount factor to its average value. In states of the world where the stochastic discount factor is greater than its average value, the pseudo probability exceeds the true probability. For example, if $m_s = \delta U'(C_{1s}) / U'(C_0)$, $\widehat{\pi}_s$ exceeds π_s in states of the world with relatively low consumption where marginal utility is high.

As a special case, suppose that in each state of nature, the stochastic discount factor equaled the risk-free discount factor; that is, $m_s = \frac{1}{R_f} = E[m]$. This circumstance implies that the pseudo probability equals the true probability and $P_a = E[mX_a] = E[X_a] / R_f$. Because the price equals the expected payoff discounted at the risk-free rate, the asset is priced as if investors are risk-neutral. Hence, this explains why $\widehat{\pi}_s$ is referred to as the *risk-neutral* probability and $\widehat{E}[\cdot]$ is referred to as the risk-neutral expectations operator. In comparison, the true probabilities, π_s, are frequently called the *physical,* or *statistical*, probabilities.

If the stochastic discount factor is interpreted as the marginal rate of substitution, then we see that $\widehat{\pi}_s$ is higher than π_s in states where the marginal utility of consumption is high (or the level of consumption is low). Thus, relative to the physical probabilities, the risk-neutral probabilities place extra probability weight on "bad" states and less probability weight on "good" states.

4.3.4 State Pricing Extensions

The complete markets pricing framework that we have just outlined is also known as *State Preference Theory* and can be generalized to an infinite number

of states and elementary securities. Basically, this is done by defining probability densities of states and replacing the summations in expressions like (4.43) and (4.44) with integrals. For example, let states be indexed by all possible points on the real line between 0 and 1; that is, the state $s \in (0, 1)$. Also let $p(s)$ be the price (density) of a primitive security that pays 1 unit in state s, 0 otherwise. Further, define $X_a(s)$ as the cashflow paid by security a in state s. Then, analogous to (4.43), we can write

$$\int_0^1 p(s)\, ds \;=\; \frac{1}{R_f} \tag{4.49}$$

and instead of (4.44), we can write the price of security a as

$$P_a \;=\; \int_0^1 p(s)\, X_a(s)\, ds \tag{4.50}$$

In some cases, namely, where markets are intertemporally complete, State Preference Theory can be extended to allow assets' cashflows to occur at different dates in the future. This generalization is sometimes referred to as Time State Preference Theory.[15] To illustrate, suppose that assets can pay cashflows at both date 1 and date 2 in the future. Let s_1 be a state at date 1 and let s_2 be a state at date 2. States at date 2 can depend on which states were reached at date 1.

For example, suppose there are two events at each date, economic recession (r) or economic expansion (boom) (b). Then we could define $s_1 \in \{r_1, b_1\}$ and $s_2 \in \{r_1r_2, r_1b_2, b_1r_2, b_1b_2\}$. By assigning suitable probabilities and primitive security state prices for assets that pay cashflows of 1 unit in each of these six states, we can sum (or integrate) over both time and states at a given date to obtain prices of complex securities. Thus, when primitive security prices exist at

[15] See Steward C. Myers (Myers 1968).

all states for all future dates, essentially we are back to a single-period complete markets framework, and the analysis is the same as that derived previously.

4.4 Summary

This chapter began by extending an individual's portfolio choice problem to include an initial consumption-savings decision. With this modification, we showed that an optimal portfolio is one where assets' expected marginal utility-weighted returns are equalized. Also, the individual's optimal level of savings involves an intertemporal trade-off where the marginal utility of current consumption is equated to the expected marginal utility of future consumption.

The individual's optimal decision rules can be reinterpreted as an asset pricing formula. This formula values assets' returns using a stochastic discount factor equal to the marginal rate of substitution between present and future consumption. Importantly, the stochastic discount factor is independent of the asset being priced and determines the asset's risk premium based on the covariance of the asset's return with the marginal utility of consumption. Moreover, this consumption-based stochastic discount factor approach places restrictions on assets' risk premia relative to the volatility of consumption. However, these restrictions appear to be violated when empirical evidence is interpreted using standard utility specifications.

This contrary empirical evidence does not automatically invalidate the stochastic discount factor approach to pricing assets. Rather than deriving discount factors as the marginal rate of substituting present for future consumption, we showed that they can be derived based on the alternative assumptions of market completeness and an absence of arbitrage. When assets' returns spanned the economy's states of nature, state prices for valuing any derivative asset could be derived. Finally, we showed how an alternative risk-neutral pric-

ing formula could be derived by transforming the states' physical probabilities to reflect an adjustment for risk. Risk-neutral pricing is an important valuation tool in many areas of asset pricing, and it will be applied frequently in future chapters.

4.5 Exercises

1. Consider the one-period model of consumption and portfolio choice. Suppose that individuals can invest in a one-period bond that pays a riskless real return of R_{rf} and in a one-period bond that pays a riskless nominal return of R_{nf}. Derive an expression for R_{rf} in terms of R_{nf}, $E[I_{01}]$, and $Cov(M_{01}, I_{01})$.

2. Assume there is an economy with k states of nature and where the following asset pricing formula holds:

$$\begin{aligned} P_a &= \sum_{s=1}^{k} \pi_s m_s X_{sa} \\ &= E[mX_a] \end{aligned}$$

   Let an individual in this economy have the utility function $\ln(C_0) + E[\delta \ln(C_1)]$, and let C_0^* be her equilibrium consumption at date 0 and C_s^* be her equilibrium consumption at date 1 in state s, $s = 1, ..., k$. Denote the date 0 price of elementary security s as p_s, and derive an expression for it in terms of the individual's equilibrium consumption.

3. Consider the one-period consumption-portfolio choice problem. The individual's first-order conditions lead to the general relationship

$$1 = E[m_{01} R_s]$$

where m_{01} is the stochastic discount factor between dates 0 and 1, and R_s is the one-period stochastic return on any security in which the individual can invest. Let there be a finite number of date 1 states where π_s is the probability of state s. Also assume markets are complete and consider the above relationship for primitive security s; that is, let R_s be the rate of return on primitive (or elementary) security s. The individual's elasticity of intertemporal substitution is defined as

$$\varepsilon^I \equiv \frac{R_s}{C_s/C_0} \frac{d\left(C_s/C_0\right)}{dR_s}$$

where C_0 is the individual's consumption at date 0 and C_s is the individual's consumption at date 1 in state s. If the individual's expected utility is given by

$$U\left(C_0\right) + \delta E\left[U\left(\widetilde{C}_1\right)\right]$$

where utility displays constant relative risk aversion, $U\left(C\right) = C^\gamma/\gamma$, solve for the elasticity of intertemporal substitution, ε^I.

4. Consider an economy with $k = 2$ states of nature, a "good" state and a "bad" state.[16] There are two assets, a risk-free asset with $R_f = 1.05$ and a second risky asset that pays cashflows

$$X_2 = \begin{bmatrix} 10 \\ 5 \end{bmatrix}$$

The current price of the risky asset is 6.

a. Solve for the prices of the elementary securities p_1 and p_2 and the risk-neutral probabilities of the two states.

[16] I thank Michael Cliff of Virginia Tech for suggesting this example.

b. Suppose that the physical probabilities of the two states are $\pi_1 = \pi_2 = 0.5$. What is the stochastic discount factor for the two states?

5. Consider a one-period economy with two end-of-period states. An option contract pays 3 in state 1 and 0 in state 2 and has a current price of 1. A forward contract pays 3 in state 1 and -2 in state 2. What are the one-period risk-free return and the risk-neutral probabilities of the two states?

6. This question asks you to relate the stochastic discount factor pricing relationship to the CAPM. The CAPM can be expressed as

$$E[R_i] = R_f + \beta_i \gamma$$

where $E[\cdot]$ is the expectation operator, R_i is the realized return on asset i, R_f is the risk-free return, β_i is asset i's beta, and γ is a positive market risk premium. Now, consider a stochastic discount factor of the form

$$m = a + bR_m$$

where a and b are constants and R_m is the realized return on the market portfolio. Also, denote the variance of the return on the market portfolio as σ_m^2.

a. Derive an expression for γ as a function of a, b, $E[R_m]$, and σ_m^2. (Hint: you may want to start from the equilibrium expression $0 = E[m(R_i - R_f)]$.)

b. Note that the equation $1 = E[mR_i]$ holds for all assets. Consider the case of the risk-free asset and the case of the market portfolio, and solve for a and b as a function of R_f, $E[R_m]$, and σ_m^2.

c. Using the formula for a and b in part (b), show that $\gamma = E[R_m] - R_f$.

7. Consider a two-factor economy with multiple risky assets and a risk-free asset whose return is denoted R_f. The economy's first factor is the return on the market portfolio, R_m, and the second factor is the return on a zero-net-investment portfolio, R_z. In other words, one can interpret the second factor as the return on a portfolio that is long one asset and short another asset, where the long and short positions are equal in magnitude (e.g., $R_z = R_a - R_b$) and where R_a and R_b are the returns on the assets that are long and short, respectively. It is assumed that $Cov(R_m, R_z) = 0$. The expected returns on all assets in the economy satisfy the APT relationship

$$E[R_i] = \lambda_0 + \beta_{im}\lambda_m + \beta_{iz}\lambda_z \qquad (*)$$

where R_i is the return on an arbitrary asset i, $\beta_{im} = Cov(R_i, R_m)/\sigma_m^2$, $\beta_{iz} = Cov(R_i, R_z)/\sigma_z^2$, and λ_m and λ_z are the risk premiums for factors 1 and 2, respectively.

Now suppose you are given the stochastic discount factor for this economy, m, measured over the same time period as the above asset returns. It is given by

$$m = a + bR_m + cR_z \qquad (**)$$

where a, b, and c are known constants. Given knowledge of this stochastic discount factor in equation (**), show how you can solve for λ_0, λ_m, and λ_z in equation (*) in terms of a, b, c, σ_m, and σ_z. Just write down the conditions that would allow you to solve for the λ_0, λ_m, and λ_z. You need not derive explicit solutions for the λ's since the conditions are nonlinear and may be tedious to manipulate.