# 1
# Consumption-Based Model and Overview

An investor must decide how much to save and how much to consume, and what portfolio of assets to hold. The most basic pricing equation comes from the first-order condition for that decision. The marginal utility loss of consuming a little less today and buying a little more of the asset should equal the marginal utility gain of consuming a little more of the asset's payoff in the future. If the price and payoff do not satisfy this relation, the investor should buy more or less of the asset. It follows that the asset's price should equal the expected discounted value of the asset's payoff, using the investor's marginal utility to discount the payoff. With this simple idea, I present many classic issues in finance.

Interest rates are related to expected marginal utility growth, and hence to the expected path of consumption. In a time of high real interest rates, it makes sense to save, buy bonds, and then consume more tomorrow. Therefore, high real interest rates should be associated with an expectation of growing consumption.

Most importantly, risk corrections to asset prices should be driven by the covariance of asset payoffs with marginal utility and hence by the covariance of asset payoffs with consumption. Other things equal, an asset that does badly in states of nature like a recession, in which the investor feels poor and is consuming little, is less desirable than an asset that does badly in states of nature like a boom in which the investor feels wealthy and is consuming a great deal. The former asset will sell for a lower price; its price will reflect a discount for its "riskiness," and this riskiness depends on a *co*-variance, not a variance.

Marginal utility, not consumption, is the fundamental measure of how you feel. Most of the theory of asset pricing is about how to go from marginal utility to observable indicators. Consumption is low when marginal utility is high, of course, so consumption may be a useful indicator. Consumption is also low and marginal utility is high when the investor's other assets have done poorly; thus we may expect that prices are low for assets that covary

positively with a large index such as the market portfolio. This is a Capital Asset Pricing Model. We will see a wide variety of additional indicators for marginal utility, things against which to compute a covariance in order to predict the risk-adjustment for prices.

## 1.1 Basic Pricing Equation

An investor's first-order conditions give the basic consumption-based model,

$$p_t = E_t\left[\beta \frac{u'(c_{t+1})}{u'(c_t)} x_{t+1}\right].$$

Our basic objective is to figure out the value of any stream of uncertain cash flows. I start with an apparently simple case, which turns out to capture very general situations.

Let us find the value at time $t$ of a *payoff* $x_{t+1}$. If you buy a stock today, the payoff next period is the stock price plus dividend, $x_{t+1} = p_{t+1} + d_{t+1}$. $x_{t+1}$ is a random variable: an investor does not know exactly how much he will get from his investment, but he can assess the probability of various possible outcomes. Do not confuse the *payoff* $x_{t+1}$ with the *profit* or *return*; $x_{t+1}$ is the value of the investment at time $t + 1$, without subtracting or dividing by the cost of the investment.

We find the value of this payoff by asking what it is worth to a typical investor. To do this, we need a convenient mathematical formalism to capture what an investor wants. We model investors by a *utility function* defined over current and future values of consumption,

$$U(c_t, c_{t+1}) = u(c_t) + \beta E_t\big[u(c_{t+1})\big],$$

where $c_t$ denotes consumption at date $t$. We often use a convenient power utility form,

$$u(c_t) = \frac{1}{1-\gamma} c_t^{1-\gamma}.$$

The limit as $\gamma \to 1$ is[1]

$$u(c) = \ln(c).$$

[1] To think about this limit precisely, add a constant to the utility function and write it as

$$u(c_t) = \frac{c_t^{1-\gamma} - 1}{1-\gamma}.$$

The utility function captures the fundamental desire for more *consumption,* rather than posit a desire for intermediate objectives such as mean and variance of portfolio returns. Consumption $c_{t+1}$ is also random; the investor does not know his wealth tomorrow, and hence how much he will decide to consume tomorrow. The period utility function $u(\cdot)$ is increasing, reflecting a desire for more consumption, and concave, reflecting the declining marginal value of additional consumption. The last bite is never as satisfying as the first.

This formalism captures investors' impatience and their aversion to risk, so we can quantitatively correct for the risk and delay of cash flows. Discounting the future by $\beta$ captures impatience, and $\beta$ is called the *subjective discount factor.* The curvature of the utility function generates aversion to risk and to intertemporal substitution: The investor prefers a consumption stream that is steady over time and across states of nature.

Now, assume that the investor can freely buy or sell as much of the payoff $x_{t+1}$ as he wishes, at a price $p_t$. How much will he buy or sell? To find the answer, denote by $e$ the original consumption level (if the investor bought none of the asset), and denote by $\xi$ the amount of the asset he chooses to buy. Then, his problem is

$$\max_{\{\xi\}} u(c_t) + E_t\left[\beta u(c_{t+1})\right] \quad s.t.$$

$$c_t = e_t - p_t\xi,$$
$$c_{t+1} = e_{t+1} + x_{t+1}\xi.$$

Substituting the constraints into the objective, and setting the derivative with respect to $\xi$ equal to zero, we obtain the first-order condition for an optimal consumption and portfolio choice,

$$p_t u'(c_t) = E_t\left[\beta u'(c_{t+1})x_{t+1}\right], \tag{1.1}$$

or

$$p_t = E_t\left[\beta \frac{u'(c_{t+1})}{u'(c_t)} x_{t+1}\right]. \tag{1.2}$$

The investor buys more or less of the asset until this first-order condition holds.

Equation (1.1) expresses the standard marginal condition for an optimum: $p_t u'(c_t)$ is the loss in utility if the investor buys another unit of the asset; $E_t\left[\beta u'(c_{t+1})x_{t+1}\right]$ is the increase in (discounted, expected) utility he obtains from the extra payoff at $t+1$. The investor continues to buy or sell the asset until the marginal loss equals the marginal gain.

Equation (1.2) is *the* central asset pricing formula. Given the payoff $x_{t+1}$ and given the investor's consumption choice $c_t, c_{t+1}$, it tells you what market price $p_t$ to expect. Its economic content is simply the first-order conditions for optimal consumption and portfolio formation. Most of the theory of asset pricing just consists of specializations and manipulations of this formula.

We have stopped short of a complete solution to the model, i.e., an expression with exogenous items on the right-hand side. We relate one endogenous variable, price, to two other endogenous variables, consumption and payoffs. One can continue to solve this model and derive the optimal consumption choice $c_t, c_{t+1}$ in terms of more fundamental givens of the model. In the model I have sketched so far, those givens are the income sequence $e_t, e_{t+1}$ and a specification of the full set of assets that the investor may buy and sell. We will in fact study such fuller solutions below. However, for many purposes one can stop short of specifying (possibly wrongly) all this extra structure, and obtain very useful predictions about asset prices from (1.2), even though consumption is an endogenous variable.

## 1.2 Marginal Rate of Substitution/Stochastic Discount Factor

We break up the basic consumption-based pricing equation into

$$p = E(mx),$$

$$m = \beta \frac{u'(c_{t+1})}{u'(c_t)},$$

where $m_{t+1}$ is the *stochastic discount factor.*

A convenient way to break up the basic pricing equation (1.2) is to define the *stochastic discount factor* $m_{t+1}$

$$m_{t+1} \equiv \beta \frac{u'(c_{t+1})}{u'(c_t)}. \tag{1.3}$$

Then, the basic pricing formula (1.2) can simply be expressed as

$$p_t = E_t(m_{t+1} x_{t+1}). \tag{1.4}$$

When it is not necessary to be explicit about time subscripts or the difference between conditional and unconditional expectation, I will suppress the subscripts and just write $p = E(mx)$. The price always comes at $t$, the payoff at $t + 1$, and the expectation is conditional on time-$t$ information.

The term *stochastic discount factor* refers to the way $m$ generalizes standard discount factor ideas. If there is no uncertainty, we can express prices via the standard present value formula

$$p_t = \frac{1}{R^f} x_{t+1}, \tag{1.5}$$

where $R^f$ is the gross risk-free rate. $1/R^f$ is the *discount factor*. Since gross interest rates are typically greater than one, the payoff $x_{t+1}$ sells "at a discount." Riskier assets have lower prices than equivalent risk-free assets, so they are often valued by using risk-adjusted discount factors,

$$p_t^i = \frac{1}{R^i} E_t\left(x_{t+1}^i\right).$$

Here, I have added the $i$ superscript to emphasize that each risky asset $i$ must be discounted by an asset-specific risk-adjusted discount factor $1/R^i$.

In this context, equation (1.4) is obviously a generalization, and it says something deep: one can incorporate all risk corrections by defining a *single* stochastic discount factor—the same one for each asset—and putting it inside the expectation. $m_{t+1}$ is *stochastic* or *random* because it is not known with certainty at time $t$. The correlation between the random components of the common discount factor $m$ and the asset-specific payoff $x^i$ generate asset-specific risk corrections.

$m_{t+1}$ is also often called the *marginal rate of substitution* after (1.3). In that equation, $m_{t+1}$ is the rate at which the investor is willing to substitute consumption at time $t+1$ for consumption at time $t$. $m_{t+1}$ is sometimes also called the *pricing kernel*. If you know what a kernel is and you express the expectation as an integral, you can see where the name comes from. It is sometimes called a *change of measure* or a *state-price density*.

For the moment, introducing the discount factor $m$ and breaking the basic pricing equation (1.2) into (1.3) and (1.4) is just a notational convenience. However, it represents a much deeper and more useful separation. For example, notice that $p = E(mx)$ would still be valid if we changed the utility function, but we would have a different function connecting $m$ to data. *All* asset pricing models amount to alternative ways of connecting the stochastic discount factor to data. At the same time, we will study lots of alternative expressions of $p = E(mx)$, and we can summarize many empirical approaches by applying them to $p = E(mx)$. By separating our models into these two components, we do not have to redo all that elaboration for each asset pricing model.

## 1.3 Prices, Payoffs, and Notation

The *price* $p_t$ gives rights to a *payoff* $x_{t+1}$. In practice, this notation covers a variety of cases, including the following:

| | Price $p_t$ | Payoff $x_{t+1}$ |
|---|---|---|
| Stock | $p_t$ | $p_{t+1} + d_{t+1}$ |
| Return | 1 | $R_{t+1}$ |
| Price-dividend ratio | $\frac{p_t}{d_t}$ | $\left(\frac{p_{t+1}}{d_{t+1}} + 1\right)\frac{d_{t+1}}{d_t}$ |
| Excess return | 0 | $R^e_{t+1} = R^a_{t+1} - R^b_{t+1}$ |
| Managed portfolio | $z_t$ | $z_t R_{t+1}$ |
| Moment condition | $E(p_t z_t)$ | $x_{t+1} z_t$ |
| One-period bond | $p_t$ | 1 |
| Risk-free rate | 1 | $R^f$ |
| Option | $C$ | $\max(S_T - K, 0)$ |

The price $p_t$ and payoff $x_{t+1}$ seem like a very restrictive kind of security. In fact, this notation is quite general and allows us easily to accommodate many different asset pricing questions. In particular, we can cover stocks, bonds, and options and make clear that there is one theory for all asset pricing.

For stocks, the one-period payoff is of course the next price plus dividend, $x_{t+1} = p_{t+1} + d_{t+1}$. We frequently divide the payoff $x_{t+1}$ by the price $p_t$ to obtain a *gross return*

$$R_{t+1} \equiv \frac{x_{t+1}}{p_t}.$$

We can think of a return as a payoff with price one. If you pay one dollar today, the return is how many dollars or units of consumption you get tomorrow. Thus, returns obey

$$1 = E(mR),$$

which is by far the most important special case of the basic formula $p = E(mx)$. I use capital letters to denote *gross* returns $R$, which have a numerical value like 1.05. I use lowercase letters to denote *net* returns $r = R - 1$ or log (continuously compounded) returns $r = \ln(R)$, both of which have numerical values like 0.05. One may also quote *percent* returns $100 \times r$.

Returns are often used in empirical work because they are typically stationary over time. (Stationary in the statistical sense; they do not have

trends and you can meaningfully take an average. "Stationary" does not mean constant.) However, thinking in terms of returns takes us away from the central task of finding asset *prices*. Dividing by dividends and creating a payoff of the form

$$x_{t+1} = \left(1 + \frac{p_{t+1}}{d_{t+1}}\right)\frac{d_{t+1}}{d_t}$$

corresponding to a price $p_t/d_t$ is a way to look at prices but still to examine stationary variables.

Not everything can be reduced to a return. If you borrow a dollar at the interest rate $R^f$ and invest it in an asset with return $R$, you pay no money out-of-pocket today, and get the payoff $R - R^f$. This is a payoff with a *zero* price, so you obviously cannot divide payoff by price to get a return. Zero price does not imply zero payoff. It is a bet in which the value of the chance of losing exactly balances the value of the chance of winning, so that no money changes hands when the bet is made. It is common to study equity strategies in which one short-sells one stock or portfolio and invests the proceeds in another stock or portfolio, generating an excess return. I denote any such difference between returns as an *excess return*, $R^e$. It is also called a *zero-cost portfolio*.

In fact, much asset pricing focuses on excess returns. Our economic understanding of interest rate variation turns out to have little to do with our understanding of risk premia, so it is convenient to separate the two phenomena by looking at interest rates and excess returns separately.

We also want to think about the *managed portfolios*, in which one invests more or less in an asset according to some signal. The "price" of such a strategy is the amount invested at time $t$, say $z_t$, and the payoff is $z_t R_{t+1}$. For example, a market timing strategy might make an investment in stocks proportional to the price-dividend ratio, investing less when prices are higher. We could represent such a strategy as a payoff using $z_t = a - b(p_t/d_t)$.

When we think about conditioning information below, we will think of objects like $z_t$ as *instruments*. Then we take an unconditional expectation of $p_t z_t = E_t(m_{t+1}x_{t+1})z_t$, yielding $E(p_t z_t) = E(m_{t+1}x_{t+1}z_t)$. We can think of this operation as creating a "security" with payoff $x_{t+1}z_t$, and "price" $E(p_t z_t)$ represented with unconditional expectations.

A one-period bond is of course a claim to a unit payoff. Bonds, options, investment projects are all examples in which it is often more useful to think of prices and payoffs rather than returns.

Prices and returns can be real (denominated in goods) or nominal (denominated in dollars); $p = E(mx)$ can refer to either case. The only difference is whether we use a real or nominal discount factor. If prices, returns, and payoffs are nominal, we should use a nominal discount factor. For example, if $p$ and $x$ denote nominal values, then we can create real

prices and payoffs to write

$$\frac{p_t}{\Pi_t} = E_t\left[\left(\beta\frac{u'(c_{t+1})}{u'(c_t)}\right)\frac{x_{t+1}}{\Pi_{t+1}}\right],$$

where $\Pi$ denotes the price level (cpi). Obviously, this is the same as defining a nominal discount factor by

$$p_t = E_t\left[\left(\beta\frac{u'(c_{t+1})}{u'(c_t)}\frac{\Pi_t}{\Pi_{t+1}}\right)x_{t+1}\right].$$

To accommodate all these cases, I will simply use the notation price $p_t$ and payoff $x_{t+1}$. These symbols can denote $0, 1$, or $z_t$ and $R_t^e, R_{t+1}$, or $z_t R_{t+1}$, respectively, according to the case. Lots of other definitions of $p$ and $x$ are useful as well.

---

## 1.4 Classic Issues in Finance

I use simple manipulations of the basic pricing equation to introduce classic issues in finance: the economics of interest rates, risk adjustments, systematic versus idiosyncratic risk, expected return-beta representations, the mean-variance frontier, the slope of the mean-variance frontier, time-varying expected returns, and present-value relations.

A few simple rearrangements and manipulations of the basic pricing equation $p = E(mx)$ give a lot of intuition and introduce some classic issues in finance, including determinants of the interest rate, risk corrections, idiosyncratic versus systematic risk, beta pricing models, and mean-variance frontiers.

### *Risk-Free Rate*

The risk-free rate is related to the discount factor by

$$R^f = 1/E(m).$$

With lognormal consumption growth and power utility,

$$r_t^f = \delta + \gamma E_t(\Delta \ln c_{t+1}) - \frac{\gamma^2}{2}\sigma_t^2(\Delta \ln c_{t+1}).$$

Real interest rates are high when people are impatient ($\delta$), when expected consumption growth is high (intertemporal substitution), or when risk is low (precautionary saving). A more curved utility function ($\gamma$) or a lower elasticity of intertemporal substitution ($1/\gamma$) means that interest rates are more sensitive to changes in expected consumption growth.

The risk-free rate is given by

$$R^f = 1/E(m). \tag{1.6}$$

The risk-free rate is known ahead of time, so $p = E(mx)$ becomes $1 = E(mR^f) = E(m)R^f$.

If a risk-free security is not traded, we can define $R^f = 1/E(m)$ as the "shadow" risk-free rate. In some models it is called the "zero-beta" rate. If one introduced a risk-free security with return $R^f = 1/E(m)$, investors would be just indifferent to buying or selling it. I use $R^f$ to simplify formulas below with this understanding.

To think about the economics behind real interest rates in a simple setup, use power utility $u'(c) = c^{-\gamma}$. Start by turning off uncertainty, in which case

$$R^f = \frac{1}{\beta}\left(\frac{c_{t+1}}{c_t}\right)^{\gamma}.$$

We can see three effects right away:

1. Real interest rates are high when people are impatient, i.e. when $\beta$ is low. If everyone wants to consume now, it takes a high interest rate to convince them to save.
2. Real interest rates are high when consumption growth is high. In times of high interest rates, it pays investors to consume less now, invest more, and consume more in the future. Thus, high interest rates lower the level of consumption today, while raising its growth rate from today to tomorrow.
3. Real interest rates are more sensitive to consumption growth if the power parameter $\gamma$ is large. If utility is highly curved, the investor cares more about maintaining a consumption profile that is smooth over time, and is less willing to rearrange consumption over time in response to interest rate incentives. Thus it takes a larger interest rate change to induce him to a given consumption growth.

To understand how interest rates behave when there is some uncertainty, I specify that consumption growth is lognormally distributed. In this case, the real risk-free rate equation becomes

$$r_t^f = \delta + \gamma E_t(\Delta \ln c_{t+1}) - \frac{\gamma^2}{2}\sigma_t^2(\Delta \ln c_{t+1}), \tag{1.7}$$

where I have defined the log risk-free rate $r_t^f$ and subjective discount rate $\delta$ by

$$r_t^f = \ln R_t^f; \qquad \beta = e^{-\delta},$$

and $\Delta$ denotes the first difference operator,

$$\Delta \ln c_{t+1} = \ln c_{t+1} - \ln c_t.$$

To derive expression (1.7) for the risk-free rate, start with

$$R_t^f = 1/E_t\left[\beta\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}\right].$$

Using the fact that normal $z$ means

$$E\left(e^z\right) = e^{E(z)+(1/2)\sigma^2(z)}$$

(you can check this by writing out the integral that defines the expectation), we have

$$R_t^f = \left[e^{-\delta}e^{-\gamma E_t(\Delta \ln c_{t+1})+(\gamma^2/2)\sigma_t^2(\Delta \ln c_{t+1})}\right]^{-1}.$$

Then take logarithms. The combination of lognormal distributions and power utility is one of the basic tricks to getting analytical solutions in this kind of model. Section 1.5 shows how to get the same result in continuous time.

Looking at (1.7), we see the same results as we had with the deterministic case. Real interest rates are high when impatience $\delta$ is high and when consumption growth is high; higher $\gamma$ makes interest rates more sensitive to consumption growth. The new $\sigma^2$ term captures *precautionary savings*. When consumption is more volatile, people with this utility function are more worried about the low consumption states than they are pleased by the high consumption states. Therefore, people want to save more, driving down interest rates.

We can also read the same terms backwards: consumption growth is high when real interest rates are high, since people save more now and spend it in the future, and consumption is less sensitive to interest rates as the desire for a smooth consumption stream, captured by $\gamma$, rises. Section 2.2 takes up the question of which way we should read this equation—as consumption determining interest rates, or as interest rates determining consumption.

For the power utility function, the curvature parameter $\gamma$ simultaneously controls intertemporal substitution—aversion to a consumption stream that varies over time, risk aversion—aversion to a consumption stream that varies across states of nature, and precautionary savings, which turns out to depend on the third derivative of the utility function. This link is particular to the power utility function. More general utility functions loosen the links between these three quantities.

*Risk Corrections*

Payoffs that are positively correlated with consumption growth have lower prices, to compensate investors for risk.

$$p = \frac{E(x)}{R^f} + \mathrm{cov}(m, x),$$
$$E(R^i) - R^f = -R^f \,\mathrm{cov}\big(m, R^i\big).$$

Expected returns are proportional to the covariance of returns with discount factors.

Using the definition of covariance $\mathrm{cov}(m, x) = E(mx) - E(m)E(x)$, we can write $p = E(mx)$ as

$$p = E(m)E(x) + \mathrm{cov}(m, x). \tag{1.8}$$

Substituting the risk-free rate equation (1.6), we obtain

$$p = \frac{E(x)}{R^f} + \mathrm{cov}(m, x). \tag{1.9}$$

The first term in (1.9) is the standard discounted present-value formula. This is the asset's price in a risk-neutral world—if consumption is constant or if utility is linear. The second term is a *risk adjustment.* An asset whose payoff covaries positively with the discount factor has its price raised and vice versa.

To understand the risk adjustment, substitute back for $m$ in terms of consumption, to obtain

$$p = \frac{E(x)}{R^f} + \frac{\mathrm{cov}\big[\beta\, u'(c_{t+1}), x_{t+1}\big]}{u'(c_t)}. \tag{1.10}$$

Marginal utility $u'(c)$ declines as $c$ rises. Thus, an asset's price is lowered if its payoff covaries positively with consumption. Conversely, an asset's price is raised if it covaries negatively with consumption.

Why? Investors do not like uncertainty about consumption. If you buy an asset whose payoff covaries positively with consumption, one that pays off well when you are already feeling wealthy, and pays off badly when you are already feeling poor, that asset will make your consumption stream more volatile. You will require a low price to induce you to buy such an asset. If you buy an asset whose payoff covaries negatively with consumption, it helps to smooth consumption and so is more valuable than its expected payoff might indicate. Insurance is an extreme example. Insurance pays off exactly when

wealth and consumption would otherwise be low—you get a check when your house burns down. For this reason, you are happy to hold insurance, even though you expect to lose money—even though the price of insurance is greater than its expected payoff discounted at the risk-free rate.

To emphasize why the *covariance* of a payoff with the discount factor rather than its *variance* determines its riskiness, keep in mind that the investor cares about the volatility of consumption. He does *not* care about the volatility of his individual assets or of his portfolio, if he can keep a steady consumption. Consider then what happens to the volatility of consumption if the investor buys a little more $\xi$ of payoff $x$. $\sigma^2(c)$ becomes

$$\sigma^2(c + \xi x) = \sigma^2(c) + 2\xi \text{ cov}(c, x) + \xi^2\sigma^2(x).$$

For small (marginal) portfolio changes, the *covariance* between consumption and payoff determines the effect of adding a bit more of each payoff on the volatility of consumption.

We use returns so often that it is worth restating the same intuition for the special case that the price is 1 and the payoff is a return. Start with the basic pricing equation for returns,

$$1 = E(mR^i).$$

I denote the return $R^i$ to emphasize that the point of the theory is to distinguish the behavior of one asset $R^i$ from another $R^j$.

The asset pricing model says that, although expected *returns* can vary across time and assets, expected *discounted* returns should always be the same, 1. Applying the covariance decomposition,

$$1 = E(m)E(R^i) + \text{cov}(m, R^i) \tag{1.11}$$

and, using $R^f = 1/E(m)$,

$$E(R^i) - R^f = -R^f \text{ cov}(m, R^i) \tag{1.12}$$

or

$$E(R^i) - R^f = -\frac{\text{cov}\left[u'(c_{t+1}), R^i_{t+1}\right]}{E\left[u'(c_{t+1})\right]}. \tag{1.13}$$

All assets have an expected return equal to the risk-free rate, plus a risk adjustment. Assets whose returns covary positively with consumption make consumption more volatile, and so must promise higher expected returns to induce investors to hold them. Conversely, assets that covary negatively with consumption, such as insurance, can offer expected rates of return that are lower than the risk-free rate, or even negative (net) expected returns.

Much of finance focuses on expected returns. We think of expected returns increasing or decreasing to clear markets; we offer intuition that "riskier" securities must offer higher expected returns to get investors to hold them, rather than saying "riskier" securities trade for lower prices so that investors will hold them. Of course, a low initial price for a given payoff corresponds to a high expected return, so this is no more than a different language for the same phenomenon.

### *Idiosyncratic Risk Does Not Affect Prices*

> Only the component of a payoff perfectly correlated with the discount factor generates an extra return. *Idiosyncratic* risk, uncorrelated with the discount factor, generates no premium.

You might think that an asset with a volatile payoff is "risky" and thus should have a large risk correction. However, if the payoff is uncorrelated with the discount factor $m$, the asset receives *no* risk correction to its price, and pays an expected return equal to the risk-free rate! In equations, if

$$\text{cov}(m, x) = 0,$$

then

$$p = \frac{E(x)}{R^f},$$

no matter how large $\sigma^2(x)$. This prediction holds even if the payoff $x$ is highly volatile and investors are highly risk averse. The reason is simple: if you buy a little bit more of such an asset, it has no first-order effect on the variance of your consumption stream.

More generally, one gets no compensation or risk adjustment for holding *idiosyncratic* risk. Only *systematic* risk generates a risk correction. To give meaning to these words, we can decompose any payoff $x$ into a part correlated with the discount factor and an idiosyncratic part uncorrelated with the discount factor by running a regression,

$$x = \text{proj}(x|m) + \varepsilon.$$

Then, the price of the residual or idiosyncratic risk $\varepsilon$ is zero, and the price of $x$ is the same as the price of its projection on $m$. The projection of $x$ on $m$ is of course that part of $x$ which is perfectly correlated with $m$. The *idiosyncratic* component of any payoff is that part uncorrelated with $m$. Thus only the systematic *part* of a payoff accounts for its price.

Projection means linear regression without a constant,

$$\text{proj}(x|m) = \frac{E(mx)}{E(m^2)}m.$$

You can verify that regression residuals are orthogonal to right-hand variables $E(m\varepsilon) = 0$ from this definition. $E(m\varepsilon) = 0$ of course means that the price of $\varepsilon$ is zero,

$$p(\text{proj}(x|m)) = p\left(\frac{E(mx)}{E(m^2)}m\right) = E\left(m^2\frac{E(mx)}{E(m^2)}\right) = E(mx) = p(x).$$

The words "systematic" and "idiosyncratic" are defined differently in different contexts, which can lead to some confusion. In this decomposition, the residuals $\varepsilon$ can be correlated with each other, though they are not correlated with the discount factor. The APT starts with a factor-analytic decomposition of the covariance of payoffs, and the word "idiosyncratic" there is reserved for the component of payoffs uncorrelated with all of the other payoffs.

### *Expected Return-Beta Representation*

We can write $p = E(mx)$ as

$$E(R^i) = R^f + \beta_{i,m}\lambda_m.$$

We can express the expected return equation (1.12), for a return $R^i$, as

$$E(R^i) = R^f + \left(\frac{\text{cov}(R^i, m)}{\text{var}(m)}\right)\left(-\frac{\text{var}(m)}{E(m)}\right) \tag{1.14}$$

or

$$E(R^i) = R^f + \beta_{i,m}\lambda_m, \tag{1.15}$$

where $\beta_{i,m}$ is the regression coefficient of the return $R^i$ on $m$. This is a *beta pricing model*. It says that each expected return should be proportional to the regression coefficient, or beta, in a regression of that return on the discount factor $m$. Notice that the coefficient $\lambda_m$ is the same for all assets $i$, while the $\beta_{i,m}$ varies from asset to asset. The $\lambda_m$ is often interpreted as the *price of risk* and the $\beta$ as the *quantity* of risk in each asset. As you can see, the price of risk $\lambda_m$ depends on the volatility of the discount factor.

Obviously, there is nothing deep about saying that expected returns are proportional to betas rather than to covariances. There is a long historical

tradition and some minor convenience in favor of betas. The betas refer to the projection of $R$ on $m$ that we studied above, so you see again a sense in which only the systematic component of risk matters.

With $m = \beta(c_{t+1}/c_t)^{-\gamma}$, we can take a Taylor approximation of equation (1.14) to express betas in terms of a more concrete variable, consumption growth, rather than marginal utility. The result, which I derive more explicitly and conveniently in the continuous-time limit (1.38) below, is

$$E(R^i) = R^f + \beta_{i,\Delta c}\lambda_{\Delta c}, \tag{1.16}$$
$$\lambda_{\Delta c} = \gamma \text{ var}(\Delta c).$$

Expected returns should increase linearly with their betas on consumption growth itself. In addition, though it is treated as a free parameter in many applications, the factor risk premium $\lambda_{\Delta c}$ is determined by risk aversion and the volatility of consumption. The more risk averse people are, or the riskier their environment, the larger an expected return premium one must pay to get investors to hold risky (high beta) assets.

### *Mean-Variance Frontier*

> All asset returns lie inside a mean-variance frontier. Assets on the frontier are perfectly correlated with each other and with the discount factor. Returns on the frontier can be generated as portfolios of any two frontier returns. We can construct a discount factor from any frontier return (except $R^f$), and an expected return-beta representation holds using any frontier return (except $R^f$) as the factor.

Asset pricing theory has focused a lot on the means and variances of asset returns. Interestingly, the set of means and variances of returns is limited. All assets priced by the discount factor $m$ must obey

$$\left|E(R^i) - R^f\right| \leq \frac{\sigma(m)}{E(m)}\sigma(R^i). \tag{1.17}$$

To derive (1.17) write for a given asset return $R^i$

$$1 = E(mR^i) = E(m)E(R^i) + \rho_{m,R^i}\sigma(R^i)\sigma(m)$$

and hence

$$E(R^i) = R^f - \rho_{m,R^i}\frac{\sigma(m)}{E(m)}\sigma(R^i). \tag{1.18}$$

Correlation coefficients cannot be greater than 1 in magnitude, leading to (1.17).

This simple calculation has many interesting and classic implications.

1. Means and variances of asset returns must lie in the wedge-shaped region illustrated in Figure 1.1. The boundary of the mean-variance region in which assets can lie is called the *mean-variance frontier*. It answers a naturally interesting question, "how much mean return can you get for a given level of variance?"
2. All returns on the frontier are perfectly correlated with the discount factor: the frontier is generated by $|\rho_{m,R^i}| = 1$. Returns on the upper part of the frontier are perfectly negatively correlated with the discount factor and hence positively correlated with consumption. They are "maximally risky" and thus get the highest expected returns. Returns on the lower part of the frontier are perfectly positively correlated with the discount factor and hence negatively correlated with consumption. They thus provide the best insurance against consumption fluctuations.
3. We can go beyond perfect correlation. Consider a payoff $m/E(m^2)$. Its price is $E(m^2)/E(m^2) = 1$, so it is a return. It is on the mean-variance frontier. Thus, if we know $m$, we can construct a mean-variance efficient return. We will expand on this theme in Chapter 5, in an explicitly incomplete market.
4. All frontier returns are also perfectly correlated with each other, since they are all perfectly correlated with the discount factor. This fact implies that we can *span* or *synthesize* any frontier return from two such returns. For example, if you pick any single frontier return $R^m$, then all frontier

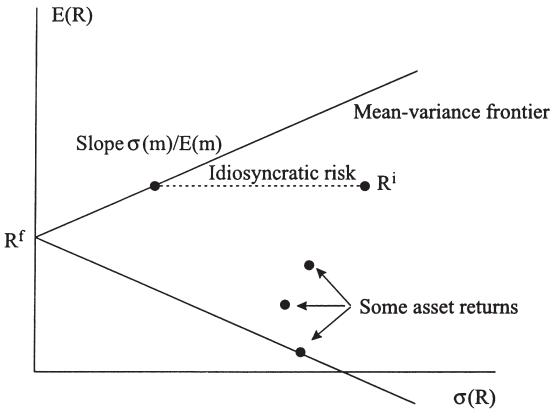


***Figure 1.1.*** *Mean-variance frontier. The mean and standard deviation of all assets priced by a discount factor m must lie in the wedge-shaped region.*

returns $R^{mv}$ must be expressible as

$$R^{mv} = R^f + a\left(R^m - R^f\right)$$

for some number $a$.

5. Since each point on the mean-variance frontier is perfectly correlated with the discount factor, we must be able to pick constants $a, b, d, e$ such that

$$m = a + bR^{mv},$$
$$R^{mv} = d + em.$$

Thus, *any mean-variance efficient return carries all pricing information*. Given a mean-variance efficient return and the risk-free rate, we can find a discount factor that prices all assets and vice versa.

6. Given a discount factor, we can also construct a single-beta representation, so *expected returns can be described in a single-beta representation using any mean-variance efficient return* (except the risk-free rate),

$$E(R^i) = R^f + \beta_{i,mv}\left[E(R^{mv}) - R^f\right].$$

The essence of the beta pricing model is that, even though the means and standard deviations of returns fill out the space inside the mean-variance frontier, a graph of mean returns versus *betas* should yield a straight line. Since the beta model applies to every return including $R^{mv}$ itself, and $R^{mv}$ has a beta of 1 on itself, we can identify the factor risk premium as $\lambda = E(R^{mv} - R^f)$.

The last two points suggest an intimate relationship between discount factors, beta models, and mean-variance frontiers. I explore this relation in detail in Chapter 6. A problem at the end of this chapter guides you through the algebra to demonstrate points 5 and 6 explicitly.

7. We can plot the decomposition of a return into a "priced" or "systematic" component and a "residual," or "idiosyncratic" component as shown in Figure 1.1. The priced part is perfectly correlated with the discount factor, and hence perfectly correlated with any frontier return. The residual or idiosyncratic part generates no expected return, so it lies flat as shown in the figure, and it is uncorrelated with the discount factor or any frontier return. Assets inside the frontier or even on the lower portion of the frontier are not "worse" than assets on the frontier. The frontier and its internal region characterize equilibrium asset returns, with rational investors happy to hold all assets. You would not want to put your whole portfolio in one "inefficient" asset, but you are happy to put some wealth in such assets.

### *Slope of the Mean-Standard Deviation Frontier and Equity Premium Puzzle*

The Sharpe ratio is limited by the volatility of the discount factor. The maximal risk-return trade-off is steeper if there is more risk or more risk aversion,

$$\left|\frac{E(R)-R^f}{\sigma(R)}\right| \leq \frac{\sigma(m)}{E(m)} \approx \gamma\sigma(\Delta \ln c).$$

This formula captures the equity premium puzzle, which suggests that either people are very risk averse, or the stock returns of the last 50 years were good luck which will not continue.

The ratio of mean excess return to standard deviation

$$\frac{E(R^i)-R^f}{\sigma(R^i)}$$

is known as the *Sharpe ratio*. It is a more interesting characterization of a security than the mean return alone. If you borrow and put more money into a security, you can increase the mean return of your position, but you do not increase the Sharpe ratio, since the standard deviation increases at the same rate as the mean.

The slope of the mean-standard deviation frontier is the largest available Sharpe ratio, and thus is naturally interesting. It answers "how much more mean return can I get by shouldering a bit more volatility in my portfolio?"

Let $R^{mv}$ denote the return of a portfolio on the frontier. From equation (1.17), the slope of the frontier is

$$\left|\frac{E(R^{mv})-R^f}{\sigma(R^{mv})}\right| = \frac{\sigma(m)}{E(m)} = \sigma(m)R^f.$$

Thus, the slope of the frontier is governed by the volatility of the discount factor.

For an economic interpretation, again consider the power utility function, $u'(c) = c^{-\gamma}$,

$$\left|\frac{E(R^{mv})-R^f}{\sigma(R^{mv})}\right| = \frac{\sigma\left[(c_{t+1}/c_t)^{-\gamma}\right]}{E\left[\left(c_{t+1}/c_t\right)^{-\gamma}\right]}. \tag{1.19}$$

The standard deviation on the right hand side is large if consumption is volatile or if $\gamma$ is large. We can state this approximation precisely using the

lognormal assumption. If consumption growth is lognormal,

$$\left|\frac{E(R^{mv}) - R^f}{\sigma(R^{mv})}\right| = \sqrt{e^{\gamma^2\sigma^2(\Delta \ln c_{t+1})} - 1} \approx \gamma\sigma(\Delta \ln c). \tag{1.20}$$

(A problem at the end of the chapter guides you through the algebra of the first equality. The relation is exact in continuous time, and thus the approximation is easiest to derive by reference to the continuous-time result; see Section 1.5.)

Reading the equation, *the slope of the mean-standard deviation frontier is higher if the economy is riskier—if consumption is more volatile—or if investors are more risk averse.* Both situations naturally make investors more reluctant to take on the extra risk of holding risky assets. Both situations also raise the slope of the expected return-beta line of the consumption beta model, (1.16). (Or, conversely, in an economy with a high Sharpe ratio, low risk-aversion investors should take on so much risk that their consumption becomes volatile.)

In postwar U.S. data, the slope of the historical mean-standard deviation frontier, or of average return-beta lines, is much higher than reasonable risk aversion and consumption volatility estimates suggest. This is the "equity premium puzzle." Over the last 50 years in the United States, real stock returns have averaged 9% with a standard deviation of about 16%, while the real return on treasury bills has been about 1%. Thus, the historical annual market Sharpe ratio has been about 0.5. Aggregate nondurable and services consumption growth had a mean and standard deviation of about 1%. We can only reconcile these facts with (1.20) if investors have a risk-aversion coefficient of 50!

Obvious ways of generalizing the calculation just make matters worse. Equation (1.20) relates consumption growth to the mean-variance frontier of all contingent claims. Market indices with 0.5 Sharpe ratios are if anything inside that frontier, so recognizing market incompleteness makes matters worse. Aggregate consumption has about 0.2 correlation with the market return, while the equality (1.20) takes the worst possible case that consumption growth and asset returns are perfectly correlated. If you add this fact, you need risk aversion of 250 to explain the market Sharpe ratio! Individuals have riskier consumption streams than aggregate, but as their risk goes up their correlation with any aggregate must decrease proportionally, so to first order recognizing individual risk will not help either.

Clearly, either (1) people are a *lot* more risk averse than we might have thought, (2) the stock returns of the last 50 years were largely good luck rather than an equilibrium compensation for risk, or (3) something is deeply wrong with the model, including the utility function and use of aggregate consumption data. This "equity premium puzzle" has attracted the attention

of a lot of research in finance, especially on the last item. I return to the equity premium in more detail in Chapter 21.

### *Random Walks and Time-Varying Expected Returns*

> If investors are risk neutral, returns are unpredictable, and prices follow martingales. In general, prices scaled by marginal utility are martingales, and returns can be predictable if investors are risk averse and if the conditional second moments of returns and discount factors vary over time. This is more plausible at long horizons.

So far, we have concentrated on the behavior of prices or expected returns across assets. We should also consider the behavior of the price or return of a given asset over time. Going back to the basic first-order condition,

$$p_t u'(c_t) = E_t[\beta u'(c_{t+1})(p_{t+1} + d_{t+1})]. \tag{1.21}$$

If investors are risk neutral, i.e., if $u(c)$ is linear or there is no variation in consumption, if the security pays no dividends between $t$ and $t+1$, and for short time horizons where $\beta$ is close to 1, this equation reduces to

$$p_t = E_t(p_{t+1}).$$

Equivalently, prices follow a time-series process of the form

$$p_{t+1} = p_t + \varepsilon_{t+1}.$$

If the variance $\sigma_t^2(\varepsilon_{t+1})$ is constant, prices follow a *random walk*. More generally, prices follow a *martingale*. Intuitively, if the price today is a lot lower than investors' expectations of the price tomorrow, then investors will try to buy the security. But this action will drive up the price of the security until the price today does equal the expected price tomorrow. Another way of saying the same thing is that returns should not be predictable; dividing by $p_t$, expected returns $E_t(p_{t+1}/p_t) = 1$ should be constant; returns should be like coin flips.

The more general equation (1.21) says that prices should follow a martingale after adjusting for dividends and scaling by marginal utility. Since martingales have useful mathematical properties, and since risk neutrality is such a simple economic environment, many asset pricing results are easily derived by scaling prices and dividends by discounted marginal utility first, and then using "risk-neutral" formulas and risk-neutral economic arguments.

Since consumption and risk aversion do not change much day to day, we might expect the random walk view to hold pretty well on a day-to-day basis. This idea contradicts the still popular notion that there are "systems" or "technical analysis" by which one can predict where stock prices are going on any given day. The random walk view has been remarkably successful. Despite decades of dredging the data, and the popularity of media reports that purport to explain where markets are going, trading rules that reliably survive transactions costs and do not implicitly expose the investor to risk have not yet been reliably demonstrated.

However, more recently, evidence has accumulated that long-horizon excess returns are quite predictable, and to some this evidence indicates that the whole enterprise of economic explanation of asset returns is flawed. To think about this issue, write our basic equation for expected returns as

$$\begin{aligned} E_t(R_{t+1}) - R_t^f &= -\frac{\mathrm{cov}_t(m_{t+1}, R_{t+1})}{E_t(m_{t+1})} \\ &= -\frac{\sigma_t(m_{t+1})}{E_t(m_{t+1})}\sigma_t(R_{t+1})\rho_t(m_{t+1}, R_{t+1}) \\ &\approx \gamma_t \sigma_t(\Delta c_{t+1})\sigma_t(R_{t+1})\rho_t(m_{t+1}, R_{t+1}), \end{aligned} \tag{1.22}$$

where $\Delta c_{t+1}$ denotes consumption growth.

I include the $t$ subscripts to emphasize that the relation applies to *conditional* moments. Sometimes, the *conditional* mean or other moment of a random variable is different from its *unconditional* moment. Conditional on tonight's weather forecast, you can better predict rain tomorrow than just knowing the average rain for that date. In the special case that random variables are i.i.d. (independent and identically distributed), like coin flips, the conditional and unconditional moments are the same, but that is a special case and not likely to be true of asset prices, returns, and macroeconomic variables. In the theory so far, we have thought of an investor, today, forming expectations of payoffs, consumption, and other variables tomorrow. Thus, the moments are really all *conditional*, and if we want to be precise we should include some notation to express this fact. I use subscripts $E_t(x_{t+1})$ to denote conditional expectation; the notation $E(x_{t+1}|I_t)$ where $I_t$ is the information set at time $t$ is more precise but a little more cumbersome.

Examining equation (1.22), we see that returns can be somewhat predictable—the expected return can vary over time. First, if the conditional variance of returns changes over time, we might expect the conditional mean return to vary as well—the return can just move in and out along a line of constant Sharpe ratio. This explanation does not seem to help much in the data; variables that forecast means do not seem to forecast variances and vice versa. Unless we want to probe the conditional correlation, predictable

excess returns have to be explained by changing risk—$\sigma_t(\Delta c_{t+1})$—or changing risk aversion $\gamma$. It is not plausible that risk or risk aversion change at daily frequencies, but fortunately returns are not predictable at daily frequencies. It is much more plausible that risk and risk aversion change over the business cycle, and this is exactly the horizon at which we see predictable excess returns. Models that make this connection precise are a very active area of current research.

*Present-Value Statement*

$$p_t = E_t \sum_{j=1}^{\infty} m_{t,\,t+j} d_{t+j}.$$

It is convenient to use only the two-period valuation, thinking of a price $p_t$ and a payoff $x_{t+1}$. But there are times when we want to relate a price to the entire cash flow stream, rather than just to one dividend and next period's price.

The most straightforward way to do this is to write out a longer-term objective,

$$E_t \sum_{j=0}^{\infty} \beta^j u(c_{t+j}).$$

Now suppose an investor can purchase a stream $\{d_{t+j}\}$ at price $p_t$. As with the two-period model, his first-order condition gives us the pricing formula directly,

$$p_t = E_t \sum_{j=1}^{\infty} \beta^j \frac{u'(c_{t+j})}{u'(c_t)} d_{t+j} = E_t \sum_{j=1}^{\infty} m_{t,\,t+j} d_{t+j}. \tag{1.23}$$

You can see that if this equation holds at time $t$ and time $t+1$, then we can derive the two-period version

$$p_t = E_t[m_{t+1}(p_{t+1} + d_{t+1})]. \tag{1.24}$$

Thus, the infinite-period and two-period models are equivalent.

(Going in the other direction is a little tougher. If you chain together (1.24), you get (1.23) plus an extra term. To get (1.23) you also need the "transversality condition" $\lim_{j\to\infty} E_t[m_{t,\,t+j} p_{t+j}] = 0$. This is an extra first-order condition of the infinite-period investor, which is not present with overlapping generations of two-period investors. It rules out

"bubbles" in which prices grow so fast that people will buy now just to resell at higher prices later, even if there are no dividends.)

From (1.23) we can write a risk adjustment to prices, as we did with one-period payoffs,

$$p_t = \sum_{j=1}^{\infty} \frac{E_t d_{t+j}}{R^f_{t,\,t+j}} + \sum_{j=1}^{\infty} \mathrm{cov}_t(d_{t+j}, m_{t,t+j}),$$

where $R^f_{t,\,t+j} \equiv E_t(m_{t,\,t+j})^{-1}$ is the $j$ period interest rate. Again, assets whose dividend streams covary negatively with marginal utility, and positively with consumption, have lower prices, since holding those assets gives the investor a more volatile consumption stream. (It is common instead to write prices as a discounted value using a risk-adjusted discount factor, e.g., $p^i_t = \sum_{j=1}^{\infty} E_t d^i_{t+j}/(R^i)^j$, but this approach is difficult to use correctly for multiperiod problems, especially when expected returns can vary over time.)

At a deeper level, the expectation in the two-period formula $p = E(mx)$ sums over states of nature. Equation (1.23) just sums over time as well and is mathematically identical.

## 1.5 Discount Factors in Continuous Time

Continuous-time versions of the basic pricing equations.

| Discrete | Continuous |
|---|---|
| $p_t = E_t \sum_{j=1}^{\infty} \beta^j \frac{u'(c_{t+j})}{u'(c_t)} D_{t+j}$ | $p_t u'(c_t) = E_t \int_{s=0}^{\infty} e^{-\delta s} u'(c_{t+s}) D_{t+s}\, ds$ |
| $m_{t+1} = \beta \frac{u'(c_{t+1})}{u'(c_t)}$ | $\Lambda_t = e^{-\delta t} u'(c_t)$ |
| $p = E(mx)$ | $0 = \Lambda D\, dt + E_t[d(\Lambda p)]$ |
| $E(R) = R^f - R^f \mathrm{cov}(m, R)$ | $E_t\left(\frac{dp}{p}\right) + \frac{D}{p}\, dt = r^f_t\, dt - E_t\left[\frac{d\Lambda}{\Lambda}\frac{dp}{p}\right]$ |

It is often convenient to express asset pricing ideas in the language of continuous-time stochastic differential equations rather than discrete-time stochastic difference equations as I have done so far. The appendix