Efficient Markets

The word efficient is too useful to be monopolized by a single meaning in economics. As a consequence, it has a variety of related but distinct meanings. In neoclassical equilibrium theory efficiency refers to Pareto efficiency. A system is Pareto efficient if there is no way to improve the well being of any one individual without making someone worse off. Productive efficiency is an implication of Pareto efficiency. An economy is productively efficient if it is not possible to produce more of any one good or service without lowering the output of some other.

In finance the word efficient has taken on quite a different meaning. A capital market is said to be (informationally) efficient if it utilizes all of the available information in setting the prices of assets. This definition is purposely vague and it is designed more to capture an intuition than to state a formal mathematical result. The basic intuition of efficient markets is that individual traders process the information that is available to them and take positions in assets in response to their information as well as to their personal situations. The market price aggregates this diverse information and in that sense it 'reflects' the available information.

The relation between the definitions of efficiency is not obvious, but it is not unreasonable to think of the efficient markets definition of finance as being a requirement for a competitive economy to be Pareto efficient. Presumably, if prices did not depend on the information available to the economy, then it would only be by accident that they could be set in such a way as to guarantee a Pareto efficient allocation (at least with respect to the commonly held information).

If the capital market is competitive and efficient, then neoclassical reasoning implies that the return that an investor expects to get on an investment in an asset will be equal to the opportunity cost of using the funds. The relation between the definitions of efficiency is not obvious, but it is not unreasonable to think of the efficient markets definition of finance as being a requirement for a competitive economy to be Pareto efficient. Presumably, if prices did not depend on the information available to the economy, then it would only be by accident that they could be set in such a way as to guarantee a Pareto efficient allocation (at least with respect to the commonly held information).

If $R_t$ denotes the total return on the asset – capital gains as well as payouts – over a holding period from $t$ to $t+1$, then the efficient markets hypothesis (EMH) asserts that

$$\mathbb{E}(R_{t+1} | I_t) = (1 + r_t),$$

(1)

where $\mathbb{E}$ is the expectation taken with respect to a given information set $I_t$ that is available at time $t$ (and that includes $r_t$). An alternative formulation of the basic EMH equation is in terms of prices. For an asset with no payouts, since

$$R_t = p_{t+1} / p_t,$$

we can rewrite (1) as

$$\mathbb{E}(p_{t+1} | I_t) = (1 + r_t)p_t,$$

or, equivalently, discounted prices must follow the martingale,

$$\frac{1}{(1 + r_t)} p_{t+1} | I_t = p_t.$$

The EMH is given empirical content by specifying the information set that is used to determine prices. Harry Roberts (1967) first coined the terms which have come to describe the categories of information sets and, concomitantly, of efficient market theories that are employed in empirical work. Fama (1970) subsequently articulated them in the form which we now use. These categories describe a hierarchy of nested information sets. As we go up the hierarchy from the smallest to the biggest set (i.e. from coarsest to finer partitions) we are requiring efficiency with respect to increasing amounts of information. At the far end of the spectrum is strong-form efficiency. Strong-form efficiency asserts that the information set, $I_t$, used by the market to set prices at each date $t$ contains all of the available information that could possibly be relevant to pricing the asset. Not only is all publicly available information embodied in the price, but all privately held information as well.

A substantial notch down from strong-form efficiency is semistrong-form efficiency. A market is efficient in the semistrong sense if it uses all of the publicly available information. The important distinction is that the information set, $I_t$, is not assumed to include privately held information, i.e. information that has not been made public. Making this distinction precise is possible in formal models but categorizing information as publicly available or not can be subjective. Presumably, accounting information such as the income statements and the balance sheets of the firm is publicly available, as is any other information that the government mandates should...
be released such as the stock holdings of the top executives in the firm. Presumably, too, the true but unrevealed intention of a major stockholder would fall into the category of private information. In between these extremes is a large grey area.

The tendency in the empirical literature has been to take a purist’s view of semistrong efficiency, and to adopt the position that if the information was in the public domain then it was available to the public and should be reflected in prices. This ignores the cost of acquiring the information, but the intuitive justification for this position is that the costs of acquiring such public information are small compared to the potential rewards. Thus, while the government mandated and publicly reported trades of the top executives require a bit more effort to obtain in a timely fashion than some average of their past holdings, such trades, when reported, would fall squarely within the realm of publicly available information under the semistrong version of the EMH.

If the asset is traded on an organized exchange, then all of the information that is clearly available to the public, none is as accessible and cheap as its past price history. At the bottom of the ladder in the efficiency hierarchy, weak-form efficiency requires only that the current and past price history be incorporated in the information set. If there is empirical validity to the EMH then, at the very least, the market for an asset should be weak-form efficient, that is, efficient with respect to its own past price history.

**Empirical testing**

The empirical implications of efficiency with respect to a particular information set are that the current price of the asset embodies all of the information in that set. Since the categories of information sets are nested, rejection of any one type, say, weak-form efficiency, implies the rejection of all stronger forms.

For example, according to weak-form efficiency, the current price of an asset embodies all of the information contained in the past price history. This implies that,

\[ E[R_t | R_{t-1}, R_{t-2}, \ldots] = (1 + r_t), \]

or, in price terms,

\[ E[p_{t+1} | p_t, p_{t-1}, \ldots] = (1 + r_t)p_t. \]

The most dramatic consequence of the EMH and certainly the one that receives the most attention from the public, is that it denies the possibility of successful trading schemes. If, for example, the market is weak-form efficient, then an investor who makes use of the ‘technical’ information of past prices can only expect to receive a return of the opportunity cost \((1 + r_t)\). No amount of clever manipulation of the past information can improve this result.

As a test of weak-form efficiency, then, we could test (although not as a simple regression) the null hypothesis that

\[ H_0 : E[p_{t+1} | p_t, p_{t-1}, \ldots] = \beta_0 + \beta_1 p_t + \beta_2 p_{t-1}, \]

where

\[ \beta_0 = 0 \]
\[ \beta_1 = (1 + r_t), \]

and

\[ \beta_2 = 0. \]

The important feature of this hypothesis is that it tells what information does not play a role (given \(r_t\)), namely the lagged price, \(p_{t-1}\). If the coefficient \(\beta_2\) should prove to be statistically significant, then this would constitute a rejection of the weak-form EMH.

The other empirical implication of the EMH that is often cited as a defining characteristic is that an efficient price series should ‘move randomly’. The precise meaning of this in our context is that price changes should be serially uncorrelated. Consider the serial covariance between two adjacent rates of return,

\[ \text{cov}(R_{t+1}, R_t) = E[(R_{t+1} - E[R_{t+1}])(R_t - E[R_t])], \]

\[ = E[R_{t+1}R_t - E[R_{t+1}]] = E[R_{t+1} | R_t] - E[R_{t+1}] \]

\[ \equiv 0. \tag{5} \]

In equation (5), since we have not specified the information set with respect to which the expectations are to be taken, they are unconditional expectations. Under weak-form efficiency, the information set will contain the past rates of return. Suppose that the (expected) opportunity cost, e.g. the interest rate \(r\), is independent of past returns on the asset or that changes are of a second order of magnitude. This would occur, for example, if we held \(r_t\) constant at \(r\). In such a case, since weak-form efficiency implies that \(H_1\) contains \(R_0\), we have

\[ E[R_{t+1} | R_t] = E[R_{t+1} | R_{t+1}] = E[R_{t+1}] \]

\[ = (1 + r_{t+1}) | R_t \]

\[ = (1 + r_t) \]

\[ = (1 + r_{t+1}). \tag{6} \]

the unconditional expectation of next period’s opportunity cost. Putting (5) and (6) together yields,

\[ \text{cov}(R_{t+1}, R_t) = E[(1 + r_{t+1})R_t - E[R_t]] = 0. \]

which is to say that rates of return are serially uncorrelated.

Tests of the EMH are legion and by and large they have been supportive. The early tests were essentially tests of the inability of trading schemes or of the random walk nature of prices, which implies that actual rates of return are serially uncorrelated. While the EMH does not imply that prices follow a random walk, such a price process is consistent with market efficiency. Alternatively, unable to specify closely the opportunity cost, some
of the early tests took refuge in the view that it must be positive, which leads to a subtranguline model for prices,

\[ E[P_{t+1} | \mathcal{F}_t] = p_t \]

The lack of a specification of the opportunity cost characterizes the early tests (see Cowles (1933), Granger and Morgenstern (1962) and Coottner (1964) and see Roll's (1984) study of the orange juice futures market for a modern example of such a test). Following Fama (1970), the literature shifted to a concern for specifying the opportunity cost and, in this sense, empirical tests became joint tests of the EMH and of the correct specification of the opportunity cost and its attendant theory.

In terms of the information hierarchy, the general message that emerged from the testing is that the market does appear to be consistent with weak-form efficiency. Tests of stronger forms of efficiency, though, have produced mixed results. Fama, Fisher, Jensen, and Roll (1969) introduced a new methodology to test semistochastic efficiency and applied it to stock splits. They observed that the residuals from a simple regression of a stock's returns on a market index would measure the portion of the price that was not attributable to market movements. By adding the residuals over a period of time, the resulting cumulative residual measures the total return over that period that is attributable to nonmarket movements. If a stock splits, say, 2 for 1, then semistochastic efficiency's price should split in proportion, i.e., half, for a 2 for 1 split. Using this 'event study' approach, Fama, Fisher, Jensen and Roll verified that stock split data was consistent with semistochastic efficiency. The event study methodology they introduced and the use of cumulative residuals (averaged over firms) has become the standard method for examining the impact of information on stock returns.

By contrast with their supportive findings, Jaffe (1974), for example, found that a rule based on the publically released information about insider trades produced abnormal returns. These results and others like them (see the section on Risk and Returns below) have been much debated and no final verdict on the matter is likely.

Recently a more interesting empirical challenge to the EMH has come from a different tack. Shiller (1981), has argued that the traditional statistical tests that have been employed are too weak to examine the EMH properly and, moreover, that they are misinterpreted. Shiller adopts the intuitive perspective that if stock prices are discounted expected dividends, then they ought not to vary over time as much as actual dividends. He argues that since the price is an expectation of the dividends and future price, what actually occurs will be this expectation plus the error in the forecast and should be more variable than the price. This leads him to formulate statistical tests of the EMH based on the volatility of stock prices which are claimed to be more powerful than the traditional (regression based) tests.

An alternative view has been taken by critics of this perspective, notably Kleidon (1986), Flavin (1983), and Marsh and Merton (1986). These critics have taken issue with Shiller's specification of the statistical tests of volatility and, more importantly, with his basic intuition. In particular, they contend that the single realization of dividends and prices that is observed is only one drawing from all of the random possibilities and that the price is based on the expectation taken over all of these possibilities. A little bit of information, then, can have an important influence on the current price. Furthermore, they argue that when the smoothing of dividends and the finite time horizon of the data samples are taken into account, volatility tests do not reject the EMH. The testing of the EMH is taking a new direction because of this work, but, at present, the results are still mixed.

Less cosmic in scope, but perhaps more worrisome is the discovery by French and Roll (1985) that the variance per unit time of market returns over periods when the market is closed (for example, from Tuesday's close to Thursday's close when the market was closed on Wednesday because of a backlog of paperwork) is many times smaller than when it is open. It is difficult to reconcile this result with the requirement that prices reflect information about the cash flows of the assets, unless the generation of fundamental information slows dramatically when the market closes – no matter why it is closed.

Theoretical formulations

The attempts to formalize the EMH as a consistent, analytical economic theory have met with less success than the empirical tests of the hypothesis. The theory can be broken into two parts. The first part is neoclassical and is largely formulated in terms of models in which investors share a common information set. Such models focus on the intertemporal aspects of the theory and the changing shape of the information set.

It has long been recognized that a competitive economy with a single risk neutral investor would lead to the traditional efficient market theories with respect to the information set employed by that investor. More interestingly, Cox, Ingersoll and Ross (1985a), and Lucas (1978) have developed intertemporal rational expectations models each of which is consistent with certain versions of the efficient market theories. There is, however, an important aspect in which these models fail to capture the essential intuition of efficient markets. In informationally efficient markets, prices communicate information to participants. Information possessed by one investor is communicated to another through the influence – however microscopic – that the first investor has on equilibrium prices. In models where investors have homogeneous information sets such information transfer is irrelevant.

A variety of attempts have been made to develop models of financial markets which can deal with such informational issues, but the task is formidable and a satisfactory resolution is not now in hand. This work parallels that of the neoclassical literature in attempting to validate the fundamental partition of the market: because of the assumption that all investors can observe the same information, the market is said to be homogeneous. In such a market, the offer to trade on the part of any one investor communicates information to other investors. In particular, it tells them that the individual, based upon his or her information, will be willing to buy the stock.

The original insight that prices reflect information about the cash flows of the assets, unless the generation of fundamental information slows dramatically when the market closes – no matter why it is closed.

To understand this point, consider a risk-averse individual trading in a market where he or she receives information signals about the ultimate value of the asset being traded and where it is common knowledge that all investors are in the same position. That is not to say that all investors have the same information, rather, it only means that they all begin with the same information, have the same view of the world (Bayesian priors), and then receive signals from the same sort of information generating mechanism. In such a market, the offer to trade on the part of any one investor communicates information to other investors. In particular, it tells them that the individual, based upon his or her information, will be willing to buy the stock. If all investors are rational they will all feel similarly bettered by trade. But, if the market had been in an equilibrium prior to the receipt of new information, and if it is common knowledge that trade balances, then in the new equilibrium not all of them can be improved. This contradiction can only be resolved by having no further trade upon the receipt of information.

To put the matter in an equivalent form, consider an investor who possesses some special information. Presumably, it is by trading that this information is incorporated into the market price. The above argument implies that the mere announcement of a wish to trade results in a change in prices with no profits for the investor since none will trade at the original prices. If information is costly to acquire and impossible to profit from, then why bother? In other words, if the price reflects the available information possessed by the individual participants, then why gather information if one only needs to look at the price?

The resolution of this dilemma can take many forms, and research will proceed by altering the assumptions that lead to this result. For example, we can drop the assumption about a common prior and let investors come to the markets with different prior beliefs. We could also drop the assumption that all investors are perfectly rational and introduce 'noisy' traders. Lastly, we could drop efficiency and complete markets to allow for insurance motives in various forms of models.

All of these approaches are being explored but we must leave this discussion with the theory that underlies the incorporation of asymmetric information into securities prices in an unsettled state. The traditional theory that prices reflect the available information is well understood with a representative individual. The theory with asymmetric information is not well understood at all. In short, the exact mechanism by which prices incorporate information is still a mystery and an attendant theory of volume is simply missing.

To conclude, the efficient market paradigm is the backbone of much of financial research and it continues to guide a large body of theoretical and empirical work. Its usefulness is beyond question, but its fine structure is not. In a sense, like much of economics, it remains a central intuition whose analytical representations seem less compelling than the insight itself. This presents more of a problem for theory than for empirical work, but the empirical results are also not without flaw. Although the evidence in support of the efficiency of capital markets is widespread, troublesome pockets of anomalies are growing and the power of the traditional methodology to test the theory is being seriously questioned. Nevertheless, there is currently no competitor for the basic intuition of efficient markets and few insights have proven as fruitful.

Risk and Return

The theory of efficient markets leads inexorably to the second central intuition in finance, the trade-off between risk and return. It has long been recognized that risk-averse investors require additional return to bear additional risk. Indeed, this insight goes back to the earliest writings on gambling and it is as much a definition of risk aversion as it is a description of risk-averse behaviour. The contribution made by finance has been to translate this observation into a body of intuition, theory, and empirics on the workings of the capital markets.

The intuition that a competitive market higher return is accompanied by higher risk owes at least as much to Calvin as it does to Adam Smith, but, in large part the development of capital market theory has been an attempt to explain risk premia, the difference between expected returns and the riskless interest rate. The foundations for the models that would first explain risk premia and that would become the workhorses of financial asset pricing theories were laid by Hicks (1946), Merton (1959), and Tobin (1958). These authors developed a rigorous micro-model of individual behaviour in a 'mean variance' world where investment portfolios were evaluated in terms of their mean returns and the total variance of their returns. They justified focusing on these two distributional characteristics by assuming either that investors had quadratic von Neumann-Morgenstern utility functions or that asset returns were normally distributed. In such a world, investors would choose mean variance efficient portfolios, i.e., portfolios with the highest mean return for a given level of variance. This observation reduced the study of portfolio choice to the analysis of the properties of the mean variance efficient set. Building on their work, Sharpe (1964),Lintner (1965), and Mossin (1966), all came to the fundamental insight that this micromodel could be aggregated into a simple model of equilibrium in the capital markets, the capital asset pricing model or CAPM.

The Mean Variance Capital Asset Pricing Model (CAPM)

In neoclassical equilibrium models, an investor evaluates an asset in terms of its marginal contribution to his or her portfolio. The decision to alter the proportion of the portfolio invested in an asset will depend on whether

the cost of doing so in terms of risk is greater or less than the benefit in expected return. An individual in a personal equilibrium will find the cost at the margin equal to the benefit.

We will assume that a unit addition of an asset to the portfolio can be financed at an interest rate of \( r \). In a mean variance model the net benefit of adding an asset to a portfolio is the additional expected return it brings, \( E_i \), less the cost of financing it. Such a change, \( Ax \), will augment the expected return on the portfolio, \( E_p \), by the risk premium of the asset, i.e. by the difference between the expected return on the asset, \( E_i \), and the cost of the financing, \( r \).

\[
AE_p = (E_i - r)Ax.
\]

The marginal cost, in terms of risk, of an increase in the holding of an asset is the addition to the total variance of the portfolio occasioned by an increase in the holding of the asset. To compute this increase, let \( v \) denote the variance of returns on the current portfolio, let \( \text{var}(i) \) stand for the variance of asset \( i \)'s returns, let \( \text{cov}(i, p) \) denote the covariance between the return of asset \( i \) and that of the portfolio, \( p \) and let \( Ax \) be the addition in the holding of asset \( i \).

The variance of the portfolio after adding \( Ax \) of asset \( i \) will be,

\[
v + Ax = v + 2Ax \text{cov}(i, p) + (Ax)^2 \text{var}(i),
\]

which means the change in the variance is given by

\[
Ax = (Ax \text{cov}(i, p) + (Ax)^2 \text{var}(i),
\]

and for a small marginal change, \( Ax \) this approximates,

\[
Ax = 2(Ax \text{cov}(i, p)).
\]

The marginal rate of transformation between return and risk, then, is given by

\[
\frac{\Delta E_p}{\Delta v} = \frac{E_i - r}{2(Ax \text{cov}(i, p)).}
\]

An investor will be in a personal equilibrium when this trade-off is equal to his or her personal marginal rate of substitution between return and risk. But, if the portfolio \( p \) is an optimal one for the investor then it must also have a trade-off between return and risk that is equal to the investor’s marginal rate of substitution, and this permits us to use it as a benchmark. Consider, then, the alternative possibility of changing the portfolio position not by changing the amount of asset \( i \) being held, but rather by changing the amount of the entire portfolio \( p \) being held, again financing the change by an alteration in the holding of the riskless asset. This is equivalent to leveraging the portfolio of risky assets and altering the amount of the riskless asset so as to continue to satisfy the budget constraint. Such a change will produce a trade-off between return and risk exactly analogous to the one examined above.

\[
\frac{\Delta E_p}{\Delta v} = \frac{E_i - r}{2(Ax \text{cov}(i, p)).}
\]

where we have written this as the marginal rate of substitution, MRS. Since in equilibrium all of the marginal rates of transformation must equal the common marginal rate of substitution, putting these two equations together we have,

\[
E_i - r = (E_p - r)\beta_p
\]

(12)

where

\[
\beta_p = \frac{\text{cov}(i, p)}{\text{var}(p)}
\]

the regression coefficient of the returns of asset \( i \) on the returns of portfolio, \( p \). Equation (12) is the famous security market line equation, the SML. It describes the necessary and sufficient condition for a portfolio \( p \) to be mean variance efficient. It also provides a clear statement of the risk premium, asserting that it is proportional to the asset’s beta, \( \beta_p \).

The insight of Sharpe, Lintner and Mossin was the observation that the SML and the mean variance analysis could be aggregated almost without change to a full equilibrium in the capital market. If we assume that all individuals have the same information and, therefore, see the same mean variance picture, then each individual’s efficient portfolio will satisfy equation (12). Since the SML equation is linear in the portfolio holding, \( p \), we can simply weight each individual’s equation by the proportion of wealth that individual holds in equilibrium, and add up the individual SML’s. The result will be an SML equation for the aggregate portfolio, \( m \), that is the weighted average of the individual portfolios. In equilibrium, the weighted average of all of the individual portfolios, \( m \), is the market portfolio, i.e., the portfolio of all assets held in proportion to their market valuation. In other words, each asset \( i \) must lie on the SML with respect to the market:

\[
E_i - r = (E_m - r)\beta_{im}
\]

which means that the market portfolio, \( m \), is a mean variance efficient portfolio.

The geometry of the mean variance analysis is illustrated in Figure 1. The set of mean variance efficient portfolios maps out a mean variance efficient frontier in the mean standard deviation space of Figure 1. Each investor will pick some point on this frontier and that point will be associated with a mean variance efficient portfolio that is suitable for the investor’s particular degree of risk aversion. All such portfolios will themselves be portfolios of just two assets: the riskless asset, \( r \), and a common portfolio, \( p \), of risky assets. This fortunate simplification of the individual portfolio optimization problem is referred to as two fund separation. It implies that the only role for individual preferences lies in choosing the appropriate combination of the risky portfolio, \( p \), and the riskless asset, \( r \). As a consequence, when we aggregate, the market risk premium, \((E_m - r)/\text{var}(m))\), will be an average of individual measures of risk aversion.

Figure 1
Black (1972) showed that two fund separation would still hold in the mean variance model even if there were no riskless asset. In such a case he found that an efficient portfolio orthogonal-the ‘zero beta portfolio’-to the market portfolio could be found, and that all investors would be able to find their optimal portfolios as combinations of m and this zero beta portfolio. In the above development of the CAPM we can simply let \( r \) be the expected return on a zero beta portfolio.

The necessary and sufficient conditions on return distributions for them to have this two fund separation property - for any concave utility function - were established by Ross (1978a). Ross characterized the class of distributions whose efficient frontier, the set of portfolios that some investor would choose, was spanned by \( k \) funds, and showed that it extended beyond the normal distribution in the case of \( k = 2 \) fund separation. This work was extended by Chamberlain (1983), who found the class of distributions for which expected utility was a function of both mean and variance for any portfolio as well as for the efficient ones. Cass and Stiglitz (1970) found the conditions on investor utility functions for a similar property to hold regardless of assumptions on return distributions.

It follows immediately from two fund separation that the tangency portfolio, \( p \), in figure 1 must be the market portfolio of risky assets since all investors hold all risky assets in the same proportions. If there is no net supply of the riskless asset then \( p \) must be the market portfolio, \( m \), itself.

The central feature of the CAPM is the mean variance efficiency of the market portfolio and the emergence of the beta coefficient on the market portfolio as the determinant of the risk premium of an asset. Those features of an asset that contribute to its variance but do not affect its covariance with the market will not influence its pricing. Only beta matters for pricing; the idiosyncratic or unsystematic risk, i.e. that portion which is the residual in the regression of the asset’s returns on the market’s returns and is therefore orthogonal to the market, playing no role in pricing.

This produces some results that were at first viewed as counter-intuitive. The older view that the risk premium depended on the asset’s variance was no longer appropriate, since if one asset had a higher variance with the market than another, it would have a higher risk premium even if the total variance of its returns were lower. Even more surprising was the implication that a risky asset that was uncorrelated with the market would have no risk premium and would be expected to have the same rate of return as the riskless asset, and that assets that were inversely correlated with the market would actually have expected returns of less than the riskless rate in equilibrium.

These results for the CAPM were supposedly explicated by the twin intuitions of diversification and systematic risk. There could be no premium for bearing unsystematic risk since a large and well diversified portfolio (i.e. one whose asset proportions are not concentrated in a small subset) would eliminate it - presumably by the law of large numbers. This would leave only systematic risk in any optimal portfolio and since this risk cannot be eliminated by diversification, it has to have a risk premium to entice risk averse investors to hold risky assets. From this perspective it becomes clear why an asset that is unrelated with the market bears no risk premium. One that is inversely correlated with the market actually offers some insurance against the all pervasive systematic risk and, therefore, there must be a payment for the insurance in the form of a negative risk premium.

There is nothing wrong with this intuition, but it does not fit the CAPM very well. The residuals from the regression of asset returns on the market portfolio are orthogonal to the market, but they could be highly correlated with each other. In fact, they are linearly independent of each other when they are weighted by the market proportions they sum to zero. This means the law of large numbers cannot be used to insure that large portfolios of residuals other than the market portfolio will be negligible. But, if that is the case, then the residuals could capture systematic risks not reflected in the market portfolio.

The CAPM was the genesis for countless empirical tests (see, e.g., Black, Jensen, and Scholes, 1972; and Fama and MacBeth, 1973). The latter paper developed the most widely used technique. The general structure of these tests was the combination of the efficient market hypothesis with time series and cross section econometrics. Typically some index of the market, such as the value weighted combination of all stocks would be chosen and a sample of firms would be tested to see if their excess returns, \( E - r \), were ‘explained’ in cross section by their betas on the index, i.e., whether the SML was rejected.

Roll (1977b, 1978) put a stop to this indiscriminate testing by calling into question precisely what was being tested. Roll’s critique had two parts. First, he argued that the tests were of very low power and probably could not detect departures from mean variance efficiency. His central point, though, began by noting that tests of the CAPM were tests of the implications of the statement that the entire market portfolio was mean variance efficient, and were not simply tests of the efficiency of some limited index such as could be formed from the stock market. The essential role played by the market portfolio in the CAPM had been stressed by others; Ross (1977b) had shown the equivalence between the CAPM and the mean variance efficiency of the market portfolio. (Ross (1976a) had also shown that in the absence of arbitrage there was always some efficient portfolio.) Roll went beyond this simple observation, though, by stressing the essential point that the market portfolio is unmeasurable. This called into question the entire cottage industry of testing the CAPM and all of the uses to which the theory had been put, such as performance measurement.

### Intertemporal models

In the aftermath of Roll’s critique, attention was turned to alternative models of asset pricing and the intertemporal nature of the theory became more important. Two separate strands of development can be traced. One essentially followed the lines of the CAPM and developed the intertemporal versions of it, the ICAPM. Merton (1973a) pioneered in this. Using continuous time diffusion analysis, Merton showed that the CAPM could be generalized to an intertemporal setting. Most interestingly, though, he demonstrated that if the economic environment was described by a finite dimensional vector of state variables, \( \xi \), a version of the SML would hold at all moments of time with the addition to the risk premium of a linear combination of the betas between the assets’ returns and each of the state variables, \( \alpha_i \).

Ross (1975) developed a similar intertemporal extension of the CAPM, but Ross’s model simplified preferences in order to close the model with an intertemporal rationality constraint and to study equilibrium price dynamics. Along the lines being developed in the modern literature on macroeconomics, intertemporal rationality and the efficient market theory required that the distribution of prices be determined endogenously. A discrete time Markov model with this feature was presented in Lucas (1978) and a full rational expectations general equilibrium in continuous time was developed in Cox, Ingersoll and Ross (1985a).

Cox, Ingersoll and Ross (1985b) applied their model to analyse and resolve some longstanding questions in the theory of the term structure of interest rates. The theory of the term structure is one of the most important subfields of finance, and the bond markets were one of the first areas where the EMH was applied. In an efficient market, ignoring risk aversion, forward rates should of future spot rates and many early theories and tests of the EMH were formulated to examine this proposition (see e.g. Mankiw, 1996). Roll (1970) integrated the EMH with the CAPM and used the resulting framework to examine empirically liquidity premia in the bond markets; the work of Cox, Ingersoll and Ross (1985b) can be considered as the logical extension of his analysis to a rational intertemporal setting.

Merton’s model was simplified markedly by Breeden (1979), who showed that if investors had intertemporally additive utility functions, then Merton’s ICAPM and its version of the SML could be collapsed back into a single beta model, the Consumption Beta model, with all assets being priced, i.e., having their risk premiums determined, by their covariance with aggregate consumption (see also Rubinstein (1976)). If we think of returns as relative prices between wealth today and in future states of nature, then optimizing individuals will set their marginal rates of substitution between consumption today and in future states equal to the rates of return. With
Arbitrage Pricing Theory (APT)

A separable but related strand of theory is the Arbitrage Pricing Theory (APT) (see e.g. Ross, 1976a, 1976b). The CAPM and the Consumption Beta model share the common feature that they explain pricing in terms of endogenous market aggregates, the market portfolio, and aggregate consumption, respectively. The APT takes a different tack.

The intuition of the CAPM (or of the Consumption Beta model) is that idiosyncratic risk can be diversified away leaving only the systematic risk to be priced. Idiosyncratic risk, though, is defined with reference to the market portfolio as the residual from a regression of returns on the market portfolio's returns. Since no further assumptions are made about the residuals, contrary to intuition a large diversified portfolio that differs from the market portfolio will not in general have insignificant residual risk. The exception is the market portfolio, but then the intuition that diversification leads to pricing by the market portfolio is circular at best.

The APT addresses this issue by assuming directly a return relationship in the systemic and idiosyncratic components of returns are defined a priori. Asset returns are assumed to satisfy a linear factor model,

$$R_i = E_i + \sum_{j} \beta_{ij} f_j + \epsilon_i, \quad i = 1, \ldots, n,$$

where $E_i$ is the expected return, $f_j$ is a demeaned exogenous factor influencing each asset $i$ through its beta on the factor, $\beta_{ij}$, and $\epsilon_i$ is an idiosyncratic mean zero term assumed to be sufficiently uncorrelated across assets that it is negligible in large portfolios. An implication of the factor structure is that the $\epsilon$ terms become negligible in large well diversified portfolios and, therefore, such portfolios approximately follow an exact factor structure,

$$R_i = E_i + \sum_{j} \beta_{ij} f_j,$$

where $i$ now denotes the $j$th well diversified portfolio. In an Arrow–Debreu state space framework, equation (16) can be interpreted as a restriction on the rank of the state-space tableaux.

An exact factor structure implies that there will be arbitrage unless the expected return on each portfolio is equal to a linear combination of the beta coefficients,

$$E_i - r = \sum_{j} \beta_{ij} f_j,$$

where $j$ is the risk premium associated with the $j$th factor, $f_j$. This equation is the APT version of the SML in the CAPM.

The APT is consistent with a wide variety of equilibrium models (including the CAPM if there is a factor structure) and it has been the object of much theoretical and empirical attention. In a sense, the APT can be thought of as a snapshot of any intertemporal model in which the factors represent innovations in the underlying state variables. This means that a rejection of the APT would imply a fairly wide ranging rejection of attempts to model asset markets with a finite set of state variables.

The original theoretical development of the APT (Ross, 1976a, 1976b) showed formally that if preferences are continuous in the quadratic mean, then the returns on a sequence of portfolios which require no wealth cannot converge to a positive return with a zero variance. This, in turn, implies that the sum of squared deviations from exact APT pricing is bounded above. These results were simplified by Huberman (1982) and extended by Ingersoll (1984) and Chamberlain and Rothschild (1983), all of whom side-stepped the issue of preferences by simply assuming that there could be no sequences converging to an arbitrage situation of a positive return with no variance. By contrast, Dybvig (1983) makes assumptions on preferences and aggregate supply to obtain a tight bound on pricing. His simple order of magnitude calculation is evidence that the pricing error is too small to be of practical significance.

By modelling the capital market explicitly as responding to innovations in exogenous variables, the APT is immediately intertemporally rational. By contrast with the CAPM and the Consumption Beta models which price assets in terms of their relation with a potentially observable and endogenous market aggregate (wealth for the CAPM and consumption for the Consumption Beta models), the APT factors are exogenous, but unspecified. Much empirical work is now underway to determine a suitable set of factors for representing systematic risk in a factor structure and to examine if they price assets successfully. (For example, see Roll and Ross, 1980; Brown and Weinstein, 1983; and Chen, Roll and Ross, 1986.)

The lack of an a priori specification for the factors has been the focus of criticism of the testability of the APT by Shanken (1982). Shanken argues that since the factors are not pre-specified, the intuitive derivation of the APT given above can be used to verify APT pricing falsely even when it does not hold, and that to prevent this some equilibrium model, such as that proposed by Connor (1984), must be used. Shanken emphasizes that his critique applies not to the theory of the APT, but rather to the way in which it has been tested. Dybvig and Ross (1985) dispute his arguments, stressing that Shanken wants to test the theory including its assumptions and approximations rather than take the positive approach of testing the model’s conclusions.

Empirical testing of asset pricing models

Since Roll’s critique, the methodology for testing asset pricing models has changed. There has been a retreat from testing a model per se to an explicit view that what is being tested is not the CAPM, for example, but rather whether the particular index being used for pricing is mean variance efficient. This change of focus has led to a more formal approach to the statistics of testing. Ross (1980) developed the maximum likelihood test statistic for the efficiency of a given portfolio and pointed out the analogy between this and the mean variance geometry, and Gibbons (1982) showed that the test of efficiency could be conducted by the use of seemingly unrelated regressions. These results have been extended by others. (For example, Kandel (1984) and Johnson and Korkie (1982)) and Gibbons, Ross and Shanken (1986) have developed and exploited an exact small sample test of the efficiency of a given index in the presence of a riskless asset. Similar tests of the APT have not yet even been developed, and to date much of the testing of the APT has focused on comparisons between the APT and pricing using the value weighted index (see e.g. Chen, Roll, and Ross, 1986).

The most important empirical finding in asset pricing, though, has been the discovery of a wide array of phenomena that appear to be consistent with nearly any neoclassical model. Consider, first, the secular effects. Asset returns fall; on average, over the weekend and rise during the week (see French, 1980). Similarly, it has been found that asset returns behave differently in the first half of the month than they do in the second. The most attention, though, has been lavished on the ‘small firm effect’. It appears that the average returns on small firms exceed those on large firms no matter what theory of asset pricing is used to correct for differences in the risk premium between these two categories of assets. Furthermore, the bulk of the return difference is concentrated in the first few days of January. Indeed, on average, returns in January appear to be abnormally large for all stocks (see e.g. Keim, 1983 or Roll, 1981, 1983).

Potentially these sorts of anomalies can be explained by secular changes in risk premia - perhaps due to secular patterns in the release of information - but their persistence and magnitude make them serious challenges to all the asset pricing models. When evidence of this sort appears difficult to explain by any pricing model it calls into question the efficient market hypothesis itself. Tests of an asset pricing model are usually joint tests of

Substitution and Arbitrage: Option Pricing

The APT is the child of one of the central intuitions of finance: namely, that close substitutes have the same price. This intuition reached fruition in the path breaking paper by Black and Scholes (1973) on option pricing. Since then the theory has found myriad applications and has been significantly extended, (see, for example, Merton (1973b), Cox and Ross (1976a, 1976b), Rubenstein (1976a, 1976b), Ingersoll (1977), Cox, Ross and Rubinstein (1979), and Cox, Ingersoll and Ross (1985a)). The Black-Scholes model employed stochastic calculus, but a simpler framework for option pricing was presented by Cox, Ross and Rubinstein (1979) that retained its essential features and was more flexible for computational purposes. We will briefly outline this binomial approach and show its connections to the major theoretical features of option pricing.

The Binomial Model

The binomial model begins with the assumption that the price of a stock, $S$, follows a proportional geometric process:
In addition to the stock there is also a riskless bond with a return of $1 + r$. The basic problem of option pricing theory is to determine the value of a derivative security, i.e., a security whose payoff depends only upon the value of an underlying primitive security, the stock in this case.

Let $C(S, t)$ denote the value of the derivative security as a function of the price of the stock and the time, $t$. Since its value depends only upon the movement of the stock – a result that is sometimes derived as a function of other attributes such as its value at the end of some period – it will also follow a binomial process:

$$C(S, t + 1) = \begin{cases} 
C(aS, t) & \text{with probability } \pi \\
C(bS, t) & \text{with probability } 1 - \pi 
\end{cases}$$

The time $t + 1$ values are illustrated in Figure 2. At any moment of time the information structure branches into relevant states, state $a$ and state $b$, defined by whether the stock goes up by $a$ or $b$. As the figure is drawn, $a > 1 + r > b$, and clearly $1 + r$ must lie between $a$ and $b$ to prevent the stock or the bond dominating. At this point there are two separate approaches to the analysis. The first is in the spirit of the original Black-Scholes model.

Suppose that at time $t$ we form a portfolio of the riskless bond and the stock with $\alpha$ dollars invested in the stock and $1 - \alpha$ dollars invested in the bond. We will choose the investment proportion so that the return on the portfolio coincides with the return on the derivative security in state $b$. This means choosing $\alpha$ so that

$$\frac{C(bS, t + 1)}{C(S, t)} = \alpha a + (1 - \alpha)(1 + r),$$

which implies that

$$\alpha = \frac{(1 + r) - C(bS, t + 1)/C(S, t)}{(1 + r) - b}.$$

But, since the portfolio’s return matches that of the derivative security in state $b$, it must also match it in state $a$. If it did not, then either the portfolio or the derivative security would dominate the other, which would be an arbitrage opportunity. In other words, we must have,

$$\frac{C(aS, t + 1)}{C(S, t)} = \alpha a + (1 - \alpha)(1 + r).$$

Putting these two equations together produces a difference equation which is satisfied by the value of the derivative security,

$$\pi C(aS, t + 1) + (1 - \pi) C(bS, t + 1) - (1 + r) C(S, t) = 0,$$

where

$$\pi S(t + 1) = \begin{cases} 
\alpha S(t) & \text{with probability } \pi \\
\beta S(t) & \text{with probability } 1 - \pi 
\end{cases}.$$
Perhaps the most remarkable feature of this equation is that it does not involve the original probabilities for the process, \( \pi \), but rather is a function of what are called the martingale probabilities, \( \pi^* \).

To solve this difference equation for the value, \( C \), of a particular derivative security we would need only to append the contractual boundary conditions that define it. For example, a European call option is specified to have the value max\( (S - E, 0) \), at a specified future date, \( T \), where \( E \) is its exercise price. Such an option gives the holder the right - but not the obligation - to buy the stock for \( E \) at time \( T \). The dual security is a European put option which gives the holder the right, but again not the obligation, to sell the stock for \( E \), at time \( T \). The problem is more difficult if the derivative security is of the American variety which means that the holder may exercise it any time up to and including the maturity date \( T \) and need not wait until \( T \).

Soon after the Black–Scholes paper, Merton (1973b) examined a variety of option contracts and showed how extensive was the range of the technique. Notably, Merton was able to derive a number of qualitative results on option pricing that were relatively independent of the particular process being modelled. For example, he showed that an American call option on a stock that pays no dividends will never be exercised before its maturity date and, therefore, will have the same value as a similar European call. He also demonstrated that put/call parity, i.e. the equivalence between the positions of holding the stock and a put option and holding a bond and a call option, was not generally valid for American options. Ross (1976c) showed that the literature’s emphasis on puts and calls was not misplaced since any derivative security could be composed of puts and calls.

A second approach to the valuation problem in our simple example illuminates why the original probabilities played no role in the analysis. Figure 2 displays what is essentially a two-state Arrow-Debreu model. In such a model if there are two pure contingent claims contracts paying one dollar in each state, then all securities can be valued as a function of their values, \( q_a \) and \( q_b \). It follows, then, that any two securities which are not linearly dependent will span the space just as two pure contingent claims would and they can be used to value all securities in the space.

In our example, the value of the bond is 1 and it must satisfy,

\[
1 = q_a (1 + r) + q_b (1 + r),
\]

and the value of the stock must satisfy,

\[
S = q_a (aS) + q_b (bS),
\]

or

\[
1 = q_a a + q_b b.
\]

Solving these two equations we can find the implicit values of the state contingent claims,

\[
q_a = \frac{(1 + r) - b}{(1 + r)(a - b)}
\]

and

\[
q_b = \frac{a - (1 + r)}{(1 + r)(a - b)}
\]

Notice that these prices do not depend on the original probability, \( \pi \), since they are derived from the values of the stock and the bond. Whatever influence the probability, \( \pi \), has on values is already reflected in the returns on the stock and the bond, and the derivative security value will just be a function of the implicit state prices. Using these prices, it is readily verified that the difference equation for the value of the derivative security, equation (23), is the same as,

\[
q_a C(aS, t + 1) + q_b C(bS, t + 1) = C(S, t).
\]

Geometrically, this means that the point,

\[
\{[C(aS, t + 1)/C(S, t)], [C(bS, t + 1)/C(S, t)]\},
\]

we plot on the same line as the return points for the bond and the stock, \( (1 + r, 1 + r) \) and \( (b, a) \). For a call option the point will be as drawn in Figure 2 indicating that the call is more volatile than the stock.

Notice from (24) and (27) that

\[
\pi^* = (1 + r) \pi_r,
\]

which means that the state space price can be interpreted as the discounted martingale probability. It is this interpretation that ties together the Cox and Ross (1976a) risk neutral approach to solving option pricing problems and the general theory of the absence of arbitrage. Cox and Ross (1976a) argued that since the difference equation that emerged for solving option pricing problems made no explicit use of any preference information, the resulting solution must also be independent of preferences. For example, then, the resulting solution must be the same as that which would obtain in a risk neutral world. In such a world, the state probabilities must be such that the expected returns on all assets are the same,

\[
\pi^* u + (1 - \pi^*) d = 1 + r,
\]
where the solution for the probability, $\pi^*$, is the same martingale probability defined above. For a European call option, then, the solution will be

$$ C(S, t) = \frac{1}{(1 + r)^T - t} \mathbb{E}^*[\max(S_T - E, 0)] $$

$$ = \frac{1}{(1 + r)^T - t} \sum_{j \geq \ln[E/S] - (T - t)/\ln(a/b)} (\pi^*)^j (1 - \pi^*)^{T - t - j} (S a_j b T - t - j - E), $$

where $\mathbb{E}^*$ is the expectation with respect to the martingale probabilities, $\pi^*$ and $(1 - \pi^*)$. It is easily verified that (29) is the solution to the difference equation (23) subject to the boundary condition,

$$ C(S, T) = \max(S - E, 0). $$

Contrast this formula with the original Black-Scholes formula for the value of a call option in a continuous time diffusion model,

$$ C(S, t) = SN(d_1) - e^{-r(T - t)} N(d_2), $$

where $N(\cdot)$ is the standard cumulative normal distribution function and,

$$ d_1 = \frac{\ln(S/E) + r(T - t) + \frac{1}{2} \sigma^2(T - t)}{\sigma \sqrt{T - t}}, $$

and

$$ d_2 = d_1 - \sigma \sqrt{T - t}. $$

Equation (30) is the solution to the Black- Scholes option pricing differential equation,

$$ \frac{1}{2} \sigma^2 S^2 C_{SS} + rSC_S - rC = -C_t, $$

subject to the boundary condition,

$$ C(S, T) = \max(S - E, 0). $$

The Black–Scholes differential equation (31) is derived from an analogous hedging argument to that for the binomial model, applied to the continuous lognormal stock process,

$$ dS/S = \mu dt + \sigma dz, $$

where $z$ is a standard Brownian motion. In fact, as the time interval between jumps converges to zero and the jump sizes shrink appropriately, the binomial converges to the lognormal diffusion and its option pricing solution will converge to that for the lognormal diffusion. Notice, too, that in analogy with the binomial whose solution does not depend upon the state probabilities, the Black-Scholes option price (30) is independent of the expected return on the stock, $\mu$.

The most interesting comparative statics result from these models is the observation that call or put option values increase with increasing variance, $\sigma^2$. This is a consequence of these options being convex functions of the terminal stock value, $S_T$ (Cox and Ross, 1976b).

The general theory of arbitrage

All of the above analysis can be tied together by the general theory of arbitrage. Under quite general conditions, it can be shown that the absence of arbitrage implies the existence of a linear pricing rule that values all of the assets (see e.g., Ross, 1976a, 1978b; Harrison and Kreps, 1979). In a static model with $m$ states of nature, this means the existence of implicit state prices, $q_j$, such that

$$ p = \sum_j q_j s_j. $$

The intertemporal extension of this result is most neatly displayed in terms of the martingale expectation used above. The absence of arbitrage now implies the existence of a martingale measure such that, with obvious notation,

$$ p = \mathbb{E}^* \left[ \exp \left( - \int_t^T r(s) ds \right) \right]. $$

This theory permits us to tie together not only the basic results of option pricing, but also our previous analysis of asset pricing models. For example, applying it to the exact factor model,
The Modigliani–Miller (henceforth MM) irrelevance propositions are developed from the absence of arbitrage, they are quite robust to alternative specifications of the economic model. To derive the Modigliani-Miller propositions we will employ the no arbitrage theory above. Consider a firm which will liquidate all of its assets at the end of the current period, and let $x$ denote the random liquidation value of the assets. Assume that the firm has debt outstanding with a face value of $F$ and that the remainder of the value of the firm is owned by the stockholders who have the residual claim after the bondholders.

At the end of the period, if $x$ is large enough the stockholders will receive $x - F$ and if $x$ falls short of $F$ they will receive nothing. Formally, then, the terminal payment to the stockholders is

$$\max(x - F, 0)$$

which will be recognized as the terminal payment on a call option. In other words – in a tribute to the ubiquitous nature of option pricing theory – the stockholders have a call option on the terminal value of the firm, $x$, with an exercise price equal to the face value of the debt, $F$. The bondholders can claim the entire assets if $x$ is not sufficient to cover the promised payment of $F$, which means that they will receive,

$$\min(x, F).$$

The current value of the firm, $V$, is defined to be the value of all of the outstanding claims against its assets which in this case is the value of the stocks, $S$, and the bonds, $B$. Using the no arbitrage analysis, we find that (ignoring discounting),

$$V = E + \sum_{j} \lambda_{j} f_{j},$$

yields the APT,

$$1 = EM(R) = EM(\left| \sum_{j} \beta_{j} f_{j} \right|) = \frac{1}{(1 + r)} \left( \sum_{j} \beta_{j} E(\left| f_{j} \right|) \right)$$

or

$$E_i = (1 + r) \sum_{j} i_{j} \beta_{ij}$$

where

$$i_{j} = -E(\left| f_{j} \right|).$$

Similarly, in a mean variance framework the martingale analysis can be used to prove that there is always a portfolio whose covariances are proportional to the excess returns on each asset. In other words, the absence of arbitrage implies the existence of a mean variance efficient portfolio (see Ross, 1976a; Chamberlain and Rothschild, 1983).

**Empirical testing**

Perhaps because the option pricing theory works so well, it has generated a surprisingly small empirical literature. Some early tests, for example, Black and Scholes (1973) and Galai (1977), focused on whether the models could be used to generate successful trading rules and found that any success was easily lost to transactions costs. Most interestingly, MacBeth and Merville (1979) found that the option formulas tended to underprice ‘in the money’ options and overprice ‘out of the money’ options, but Geske and Roll (1984) have argued that this effect disappears with a reformulation of the statistics.

Given a theory that works so well, the best empirical work will be to use it as a tool rather than to test it. Chinas and Manaster (1978), for example, show that implicit volatilities, i.e. variances computed by inverting the option formulas to obtain variance as a function of the quoted option price, have strong predictive power for explaining future realized stock variances. Patell and Wolfson (1979) use the implicit variances to examine whether stock prices are more volatile around earnings announcements.

These efforts should increase; options and option pricing theory give us an opportunity to measure directly the degree of anticipated uncertainty in the markets. Financial press terms such as ‘investor confidence’ take on new meaning when they can actually be measured.

This does not mean, however, that there are no important gaps in the theory. Perhaps of most importance, beyond numerical results (see, for example, Parkinson, 1977; or Brennan and Schwartz, 1977), very little is known about most American options which expire in finite time. The American call option on a stock paying a dividend or the American put option are both easily solved in the infinite maturity case since the optimal exercise boundary is a fixed stock value independent of time (Merton, 1973b; Cox and Ross, 1976a, 1976b). If dividends occur at discrete points, then if the call is exercised prematurely it will only be optimal to do so just prior to a dividend payment. This permits a recursive approach to the solution of this finite maturity option (see Roll, 1977a; Geske, 1979). But, with continuous payouts, surprisingly little is known about the exercise properties of either of these options in the American case.

Despite such gaps, when judged by its ability to explain the empirical data, option pricing theory is the most successful theory not only in finance, but in all of economics. It is now widely employed by the financial industry and its impact on economics has been far ranging. At a theoretical level, we now understand that option pricing theory is a manifestation of the force of arbitrage and that this is the same force that underlies much of neoclassical finance.

**The Whole is the Sum of the Parts – Corporate Finance**

The use of arbitrage as a serious tool of analysis coincided with the beginning of the modern theory of corporate finance. In two seminal papers on the cost of capital, Modigliani and Miller (1958, 1963) argued that the overall cost of capital and, therefore, the value of the firm would be unaffected by its financing decision. Specifically, using arbitrage arguments, Modigliani and Miller showed that the debt/equity split would not alter a firm’s value and they then argued that with the investment decision held constant, the dividend payout rate of the firm would also not affect that value. These two irrelevance propositions defined the study of corporate finance in much the same way that Arrow’s Impossibility Theorem defined social choice theory. At one and the same time they propounded an irreverent theory whose central feature was the irrelevance of the topic under study. This challenge, to weaken in a useful way the assumptions of their analysis, has guided research in this area ever since.

**The Modigliani–Miller analysis**

Since the Modigliani-Miller irrelevance propositions are developed from the absence of arbitrage, they are quite robust to alternative specifications of the economic model. To derive the Modigliani-Miller propositions we will employ the no arbitrage theory above. Consider a firm which will liquidate all of its assets at the end of the current period, and let $x$ denote the random liquidation value of the assets. Assume that the firm has debt outstanding with a face value of $F$ and that the remainder of the value of the firm is owned by the stockholders who have the residual claim after the bondholders.

At the end of the period, if $x$ is large enough the stockholders will receive $x - F$ and if $x$ falls short of $F$ they will receive nothing. Formally, then, the terminal payment to the stockholders is

$$\max(x - F, 0)$$

which will be recognized as the terminal payment on a call option. In other words – in a tribute to the ubiquitous nature of option pricing theory – the stockholders have a call option on the terminal value of the firm, $x$, with an exercise price equal to the face value of the debt, $F$. The bondholders can claim the entire assets if $x$ is not sufficient to cover the promised payment of $F$, which means that they will receive,

$$\min(x, F).$$

The current value of the firm, $V$, is defined to be the value of all of the outstanding claims against its assets which in this case is the value of the stocks, $S$, and the bonds, $B$. Using the no arbitrage analysis, we find that (ignoring discounting),
\[ Y = S + B \]
\[ = E[\max(x - F, 0)] + E[\min(x, F)] \]
\[ = E[\max(x - F, 0) + \text{max}(x, F)] \]
\[ = \Omega(x), \]

which is independent of the face value, \( F \), of the debt and, therefore, independent of the relative amounts of debt and equity. This verifies the first of the MM irrelevance propositions.

To verify the irrelevance of value to the dividend payout, consider a firm about to pay a dividend, \( D \). The current, pre-dividend, value of the stock is \( p(D) \) and by the no arbitrage martingale analysis this is given by,

\[ p(D) = D + E[p'(D)], \]

where \( p'(D) \) is the ex-dividend price. If the investment policy of the firm has been fixed, then the only impact that the current dividend payout can have on the stockholders is through its alteration of the cash in the firm. This means that changing the dividend to, say \( D + \Delta D \), would necessitate a change in current assets of \(-\Delta D\). From the first MM proposition the mode of financing this change in the dividend will be irrelevant to the determination of the firm’s value and to simplify the analysis we will assume that it is financed by riskless debt. At an interest rate of \( r \) this would entail, say, a perpetual outflow from the firm of \( rD \). Again applying the analysis and letting \( x_{t+s} \) be the cash flow at time \( t+s \) given that a dividend of \( D \) is paid now, we have,

\[ p'(D + \Delta D) = E\left[ \sum_{t=0}^{\infty} e^{-rt} x_{t+s} (D + \Delta D) \right] - E\left[ \sum_{t=0}^{\infty} e^{-rt} x_{t+s} D \right] \]
\[ = p'(D) - \Delta D. \]

Thus, we have the irrelevance proposition,

\[ p'(D + \Delta D) = D + E[p'(D + \Delta D)] \]
\[ = D + \Delta D + E[p'(D + \Delta D)] \]
\[ = D + E[p'(D)] - \Delta D \]
\[ = D + E[p'(D)] \]
\[ = p'(D). \]

The MM results were startling to those who had worked in corporate finance and had taken it for granted that the way in which a firm was financed affected its value. To understand the importance of the MM results for the most practical of problems; recall that the original impetus for the study of corporate finance was the determination of the firm’s opportunity cost for investments, \( p \). For a marginal investment, financed by the issuance of debt and equity, the cost of capital, \( r \), also known as the weighted average cost of capital, WACC, would be the weighted average cost of the debt, \( r \), and the cost of equity, \( k \).

\[ p = (S / V)k + (B / V)r \]

(33)

(where we have ignored tax effects).

If debt is riskless, then \( r \) is the interest rate on such debt and \( k \), the cost of equity, will be the return required by investors for the risk inherent in the stock. Presumably \( k \) could be found by appeal to one of the asset pricing models discussed above.

Now it is tempting to think, for example, that if \( k > r \), then an increase in debt relative to equity will lower \( p \). If this goes too far, debt will become risky and as \( r \) rises there will be a unique optimal debt/equity ratio, \( (B/S)^* \), that minimizes the cost of capital, \( p \). This would be the discount rate to use for present value calculations and it would maximize the value of the firm. This was the traditional analysis of the leverage decision before MM.

By the MM theorem, though, value, \( V \), is unaffected by leverage. This means that \( p \) is unchanged, since the total (expected) return to the stockholders and the bondholders, \( Sk + Br \), is unaltered [see Equation (33)]. In terms of the WACC, then, as the leverage \( (B/S) \) is increased by the substitution of debt for equity, the cost of equity changes.

\[ k = p + (B / S)(p - r), \]

but not the WACC.

Spanning arguments

The efforts to elude these results and to develop a meaningful theory of corporate finance have taken many forms. First, it has been argued that the analysis itself contains a hidden and critical assumption, namely that the pricing operator is independent of the corporate financial structure. The alternative is that the change in the debt/equity decision, for example, will also change the span of the marketed assets in the economy and, consequently, the operator used for pricing will change. The simplest such example would be a single firm in a two-state world. If the firm is an all-equity firm and if there are no other traded assets, then individuals cannot adjust their consumption across the states of nature and must split it according to the equity payoff. If this firm now issues debt the two securities will span the two states of nature and complete the market. This, in turn, will generally alter pricing in the economy.

While this argument has generated a large literature, the problem of the determination of the corporate financial structure and the value of the firm is primarily a microeconomic question and it is difficult to believe that it will be resolved or even illuminated by assuming that firms have some monopoly power that enables them to alter pricing in the capital markets. At the microlevel the MM propositions are unlikely to be seriously affected by such general equilibrium arguments.

At the microlevel, too, the intuition behind the MM propositions and its conclusions is so robust as to be daunting. Consider the following argument. According to MM there can be no optimal, i.e. value maximizing, financial structure since value is independent of structure. Suppose that there was an optimal, say, debt/equity ratio, \( (B/S)^* \). Any departure from this target \( (B/S)^* \), however, could not lower the firm’s value since it would immediately afford an arbitrage opportunity to buy the total firm at its lowered value and refinance it in the optimal target propositions, \( (B/S)^* \). (This somewhat facetious argument gets the point across, but it really means that we have not fully specified the rules of the game, e.g., who moves first, what happens when no one moves, etc.)

Signalling models

A more promising route which formally exploits incomplete spanning, but does not argue that the pricing operator itself is altered by any one firm changing its financial structure, makes use of the theory of asymmetric
information and signalling (see Ross, 1977a; Leland and Pyle, 1977; and Bhattacharya, 1979). If the managers of the firm possess information that is not held by the market then the market will make inferences from the actions of the firm and in particular, from financial decisions. Changes in its financial structure or its dividend policy will alter investors' perceptions of its risk class and, therefore, its value. While the operator, $E^{*}$, does not change, the perception of the distribution of the firm's cash flows does. An effort to maximize their value, firms will take actions, such as taking on high debt to equity ratios, which can be initiated by lesser firms only at a prohibitive cost. This will distinguish them from lesser firms that the uninformed market erroneously puts into the same class with them. In this fashion, a hierarchy of firm risk classes will emerge, and, in equilibrium, firms will signal their true situations and investors will draw correct inferences from their signals.

All of this has a nice ring to it, but the nagging question that remains is why firms use their financial decisions to accomplish all of this information transfer. Financial changes are cheap, but even clever might be guarantees or, for that matter, a system of legislation. These issues remain unresolved, but it is difficult to think that much will be explained by theories that argue that firms take on more debt just to show the world that 'they can do it'. There is a limit to mache-finance.

### Taxes

Another line of attack has been to introduce more 'imperfections', especially taxes, into the models. Modigliani and Miller originally had noted that the presence of a corporate tax means that firms would have an incentive to issue additional debt. Since interest payments on debt are excluded from corporate taxes, substituting debt for equity permits firms to pass returns to investors with a lowered tax cut to the government. At the limit, firms would be all debt if the tax authorities still recognized such debt payments as excludable from taxable corporate income. Presumably, the only brake to this expansion would be the real costs of dealing with the inevitable bankruptcies of high debt firms. This is logically possible, but at the expense of reducing corporate finance to the study of the tradeoff between the tax advantages of debt and the costs of bankruptcy.

Miller (1977) found a more profound brake to this tendency to increase debt. He argued that while the firm could lower its taxes by increasing its debt, the ability of investors to defer or offset capital gains implies that they pay higher taxes on interest income than on the returns from equities. With a rising tax, an equilibrium is possible in which the marginal investor has a tax differential between ordinary income and equity returns that exactly offsets the firm's corporate tax advantage to debt. In such an equilibrium, investors in a higher tax bracket than the marginal investor would purchase only equity (or non-taxed bonds such as municipals for US investors) and those in lower tax brackets would purchase corporate bonds. There would be an equilibrium amount of debt for the corporate sector as a whole, but not for any individual firm (assuming the absence of inframarginal firm tax schedules).

Miller's analysis led to a large literature on the impact of taxes on pricing. Black and Scholes (1974) had made a related argument for the absence of a tax effect on dividends, arguing that stocks with relatively higher yields should not have higher gross returns to compensate investors for the additional tax burden since companies would then cut their dividends to increase the stock price. Black and Scholes verified their results empirically, but, using a different methodology, Litzenberger and Ramaswamy (1982) found that gross returns were higher for stocks with higher dividends. Whether the supply side or the demand side dominates remains undecided.

Whatever the resolution of this and similar issues, the equilibrium argument initiated by Miller has changed much of the analysis of these issues. Miller and Scholes (1978), for example, argue that by employing a number of 'laundring' devices individuals can dramatically cut their taxes. Their conclusion that, in theory, taxes should be much lower than they appear to be in practice, focuses attention on the role played by informational asymmetries and the related costliness of using techniques such as investing through tax exempt intermediaries.

### Agency models

The emphasis on informational asymmetries has been the cornerstone of an alternative approach to corporate finance, agency theory. Wilson (1968) and Ross (1973) developed agency models in which one party, the agent (e.g. a corporate manager) acts on behalf of another the principal (e.g. stockholders). Jensen and Meckling (1976), building on the agency theory and on Williamson's (1975) transaction cost approach, argue that corporate finance can be understood in terms of the monitoring and bonding costs imposed on stockholders and managers by such relations. The manager qua employee has an incentive to divert firm resources to his own benefit.

Jensen and Meckling refer to the loss in value in restraining this incentive as the (equilibrium) agency cost of the relation.

To some extent this conflict can be resolved ex ante by the indenture agreements and covenants in financial contracts, but the cost of doing so rises with the monitoring requirements. Myers (1977), for example, has studied the implications for investment policy of the conflict between the stockholders and the bondholders. Stockholders own a call option on the assets of the firm and the value of a call increases with the variance of the asset value. Conversely, such increases will come at the expense of the bondholders. Ex ante indenture agreements can limit the ability of management and stockholders to take on additional risk, but the more precise the limits the costlier it is to observe, observe and enforce them.

These trade-offs are the intuition and subject matter of the agency approach to corporate finance, but to date it is more a collection of intuitions than a well-articulated theory. The agency approach has pointed in some intriguing directions, but it fares poorly if judged by asking what it is that would be a counter observation or count as evidence against it. To the contrary, no phenomenon seems beyond the reach of "agency costs" and at times the phrase takes on more of the trappings of an incantation than an analytical tool. The role of asymmetric information in corporate finance and in explaining the managerial and financial forces at work in the firm is self evident, but it remains fertile ground for theory.

### Empirical evidence

The early empirical work examined the relation between the corporate financial structure and other characteristics of the firm. Hamada (1972), for example, studied whether the beta of a firm's equity was related to the beta of the firm's assets as would be predicted by the cost of capital, equation (33). There continues to be empirical work on these issues, but the attention of empiricists has shifted to the arena of corporate control.

A boom in merger and acquisition activity in the late 1970s and through to the present time has brought some striking and unexplained empirical regularities. On average, shareholders in firms that are the targets of tender offers gain significantly from such offers while the rewards to bidders are still ambiguous (Jensen and Ruback, 1983). For unsuccessful tenders the target firms appear to average an eventual loss and the bidders may, too. These results and the discrepancy between targets and bidders have been the object of close scrutiny.

If firms realize such abnormal gains as targets, and if it reflects the release of information about the value of their underlying assets, then that raises the question of why they were not priced correctly to begin with. On the other hand, if the returns for successful targets reflect synergies rather than simply a revaluation of their assets, why does the bidder get so little? Several game theoretic and bidding models have been built in an attempt to explain these results (see, e.g., Grossman and Hart, 1980), but a consensus has yet to emerge. Furthermore, some of the important empirical issues, such as whether bidders actually gain or lose on average remain unresolved.

### Conclusion

For corporate finance, like the other major areas of finance, the neoclassical theory is now well established, but, like the other areas, the inadequacy of the neoclassical analysis is pushing researchers to begin the challenging task of explaining these results (see, e.g., Grossman and Hart, 1980), but a consensus has yet to emerge. Furthermore, some of the important empirical issues, such as whether bidders actually gain or lose on average remain unresolved.

### See Also

- arbitrage
- capital asset pricing model
- dividend policy
- efficient market hypothesis
- options

### Bibliography


**How to cite this article**