

A View of the Current Status of the Size Anomaly¹

Jonathan Berk
School of Business Administration
University of Washington

July 11, 1996

¹In preparation for Keim, D. and W. Ziemba, editors, *Security Market Imperfections and World Wide Equity Markets*, Cambridge University Press

It is in many ways fitting that a conference devoted to security market anomalies should take place at the Isaac Newton Institute. The paradigm in physics for which Isaac Newton is responsible is arguably the most influential paradigm ever conceived. Most of the technological advances since that time would not have been possible without the paradigm. Yet, there have always been puzzles that, at the time of their discovery, appeared to be inconsistent with this paradigm. For instance, in the sixty odd years that it took to reconcile the orbit of the moon with Newton's inverse square law, some scientists suggested that the law needed to be modified. Ultimately, like countless other puzzles, the orbit of the Moon was shown to be predicted perfectly by the paradigm. The paradigm is, nevertheless, wrong. The reason we know this is that not every puzzle could be reconciled with the theory — a few were true anomalies, events that could only be explained by a better paradigm. As useful as Newton's laws are, they cannot explain the motion of Mercury's perihelion while Einstein's theory of general relativity can. Without the existence of this and one or two other anomalies, it is likely that Newton's laws would still be the currently accepted paradigm (and this institute would not have hosted the Hawking-Penrose lectures on black holes.)

Although the case of Newton's law is only one example, there are good reasons for believing that puzzles and anomalies play a critical role in the advancement of knowledge. Thomas Kuhn (1970) in his seminal work on the subject, convincingly argues that the advancement of knowledge begins with the development of a paradigm. Puzzles that, at least initially, are not obviously explained by the paradigm are then identified. The field then advances by demonstrating why these puzzles are consistent with the paradigm. Sometimes, this effort fails and it becomes clear (usually only *ex post*), that these puzzles cannot be explained by the prevailing paradigm. A new paradigm that can explain these puzzles then replaces the old paradigm. The puzzles that prompted the development of the new paradigm are then known as anomalies.

In the last decade or so a number of empirical regularities in financial economics have been identified that have become known as "anomalies." Given the critical role anomalies play in the advancement of knowledge, identifying them is a significant milestone. In light of this, the question of whether these empirical regularities should legitimately be labelled as anomalies is an important one. That is, are these regularities so incongruous that a completely new asset pricing paradigm is required to explain them, or are they merely puzzles to be solved, under the existing paradigm, in the course of what Kuhn describes as normal science? In this chapter I will critically examine this question in the context of what perhaps is the most notorious empirical regularity — the size

effect.

I will present the view that the so called “size effect” (the relation between realized returns and variables such as market value, book-to-market, earnings-to-price, etc.) should not even be classed as a puzzle, let alone an anomaly. Furthermore, while I will argue that a derivative of this research, the development of sized-based factor models, might be useful in helping to explain the cross-sectional variation of expected returns, such “models” provide no information on the underlying economic cause of this variation. Finally, I will argue that the size variables nevertheless have a useful role to play in tests of asset pricing theories.

The rest of this chapter is organized as follows. In the next section I will briefly review the current empirical research on the size-related anomalies and then will try to define precisely what asset pricing paradigm researchers have in mind when they refer to this evidence as an anomaly. In Section 2, I pose the question: Is the size effect an anomaly? I will argue that it is not. Following that, in Section 3, I will examine the importance of size based factor models. In Section 4 I will consider the implication, on current asset pricing paradigms and tests of those paradigms, of the fact that both market value and book-to-market have been shown to have additional explanatory power over and above the single beta model. Section 5 concludes the chapter.

1 The Size Effect

The cross-sectional relation between market value and return (the “size anomaly”) was discovered by Banz (1981). He not only found that market value is an excellent predictor of expected return, but that it is better in this regard than the CAPM itself. Reinganum (1981) then established that another effect, the predictive power of the earnings-to-price ratio (E/P), was in fact related to the size anomaly. Indeed, Reinganum showed that once size is controlled for, the E/P anomaly disappears.¹ Stattman (1980) and Rosenberg, Reid and Lanstein (1985) document another anomaly, that the ratio of book equity to market equity (book-to-market) is a predictor of average return. Other studies include Chan, Hamao and Lakonishok (1991) who show that these anomalies are also present in the Japanese stock market and Keim (1983) and Keim and Stambaugh (1984) who document important seasonals in the empirical relation.² Recently, both book-to-market and

¹Banz (1983), however, argues that the P/E ratio still has additional explanatory power once size and book-to-market have been controlled for.

²The reader who is interested in a more comprehensive survey is directed to one of the many excellent texts that document the various stock market anomalies (e.g., Dimson (1988) or Ziemba (1994).)

market value have also been shown to provide predictive power in the time-series of expected returns (see Pontiff and Schall (1995) and Kothari and Shanken (1995)).

The recent interest in the anomaly was sparked by Fama and French (1992) who systematically redocument the relation between market value (book-to-market) and return. The paper then convincingly argues that the other related affects as well as beta (from the single beta model) are all subsumed by book-to-market and market value. In Fama and French (1993) and Fama and French (1995), the authors form portfolios based on these two variables and show that their portfolios can be used to predict return in much the same way as the market portfolio is used in the CAPM. As a result they interpret these portfolios as factor mimicking portfolios and advocate using them to price assets. The authors attribute the predictive power of these portfolios to their ability to capture risk. Whether or not such an inference follows from their empirical work is the focus of Section 3.

Although the size related regularities are widely referred to as anomalies, the precise paradigm for which they present an anomaly is not so clear. Authors such as Lee, Shleifer and Thaler (1991) seem to have one of the central tenants in financial economics in mind — that the expected of an asset is determined solely by the asset’s riskiness. Other authors, such as Fama and French (1995), interpret the anomaly more narrowly. They take specific asset pricing models such as the CAPM or APT as the paradigm. In the next section I will examine the implications of the size related regularities on both of these definitions of the underlying paradigm.

2 Why the Size Effect Could Never be an Anomaly

To be classed as an anomaly the size effect must have the feature that it cannot be explained by the paradigm under consideration. Yet, as we will see, under both the narrow and broad definition of the paradigm, market value (or book-to-market) *must be* inversely related to return and so there is no sense in which the discovery of such a relation should be regarded as an anomaly.

Consider a one period economy in which expected value of every firm’s end-of-period cashflow is the same. In order to ensure that stock prices can be compared easily, also assume that each firm issues only one share of stock. Now consider populating the world with risk neutral agents. Then it follows that stock prices in this world will be equal and so the market value of all firms will be the same. Now consider the same world populated with risk averse agents. Since the riskiness of a firm’s cashflow is likely to vary across firms, each firm’s stock price will, in general, be different.

Given that all firms have the same expected cashflow, riskier firms with lower market values will, by definition, have higher expected returns. In the cross-section, market value will be inversely related to expected return. Finally, consider the possibility that stock prices are not influenced by the risk preferences of agents. For the sake of argument, assume that green men on Mars pick stock prices. Since all stocks have the same expected cashflow, by definition, the stocks that these green men assign lower (higher) prices will have higher (lower) expected returns. Thus, even in this world, market value is cross-sectionally inversely related to return. Clearly the only condition under which the size effect will not be observed in such a world is if all assets have the same price. *Any* theory that successfully explains any cross-sectional variation in expected returns *must* also predict an inverse relation between expected return and market value. In such a world, the size effect could never be an anomaly.

One shortcoming of the above argument is the assumption that all firms have the same expected cashflow. However, as I have shown elsewhere (Berk (1995)), so long as expected returns are not positively correlated to expected cashflows, the above result extends to an economy in which expected cashflows are not equal. This argument therefore shows that the only asset pricing paradigms for which the size effect could be an anomaly are ones which require a *positive* cross-sectional correlation between expected returns and expected cashflows. Since I know of no asset pricing theory that has this requirement, the size effect cannot be inconsistent with any paradigm that I am aware of. In particular it is neither inconsistent with the hypothesis that risk and return are related nor is it inconsistent with a specific asset pricing model such as the CAPM or APT.

Finally, consider what would happen if, instead of investigating the relation between expected return and market value, the relation between expected return and the ratio of expected cashflows to market value is investigated. In a single period model, another name for the ratio of expected cashflows to market value is the expected return, so this ratio is tautologically perfectly correlated to expected return! This observation requires no auxiliary assumptions whatsoever. Unfortunately, expected cashflows are unobservable and so no such study has been undertaken. However, something quite similar has been done.

The book value of equity measures depreciated past investment. Since the amount invested is likely to be highly correlated with the expected cashflows of the investment, one would expect the book value of equity and expected cashflow to be highly correlated. Thus book equity can be used

as a proxy for the expected cashflow.³ Therefore, the ratio of book equity to market equity is, in principle, a proxy for the ratio of expected cashflows to market value and should therefore be a better measure of expected return than market equity alone. In light of the above argument it is not surprising then, that both Fama and French (1992) and Jegadeesh (1992) find that the logarithm of book-to-market equity is a much better predictor of return than the logarithm of market equity alone.

3 What Do We Learn From Size Based Factor Models?

The fact that the relation between market value (book-to-market) and return should not be regarded as anomalous does not imply that it is not potentially useful. Consequently, Fama and French (1995) have used portfolios constructed on the basis of market value and book-to-market to construct a “factor model” of asset returns. The authors argue, based on the ability of these portfolios to explain cross-sectional variation in asset returns, that they collectively span the risk factors in the economy. In fact a portfolio that explains cross-sectional differences in asset returns need not necessarily capture risk in any form.

There are at least two reasons why a portfolio might explain cross-sectional differences in expected returns that are not risk related. The first is explained in a recent paper by Ferson (1995). In this paper, the author shows how an arbitrary set of firm specific attributes that are assumed to be related to expected returns but unrelated to any economy wide risk factor can nevertheless be used to form portfolios that will explain cross-sectional differences in expected returns. The second reason follows from the fact that the existence of a mean-variance efficient frontier does not depend on their being a risk-return trade-off in the economy. By the Roll (1977) critique, a mean-variance efficient portfolio always exists. This portfolio will always explain cross-section variation in expected return, yet its existence does not in any way depend on whether this cross-sectional variation is risk based.

Of course, the mere existence of frontier portfolios does not necessarily explain why the particular portfolios Fama and French identify work as well as they do. However, mean-variance efficient portfolios are not the only portfolios that can explain cross-sectional differences in expected return. As explained in Green (1986) even portfolios that are not mean-variance efficient will still explain

³Berk, Green and Naik (1996) provide a formal model in which value of book equity is a perfect proxy for the expected cashflow.

cross-sectional differences in asset returns. Like the Roll critique this result does not rely on a presumed relation between risk and return. Indeed, in an economy in which risk is assumed to be completely unrelated to return it is possible to derive an example in which *any* portfolio (that does not require short selling) will explain cross-sectional variation in asset returns. To see why, consider the following economy.

Take a multiperiod economy that consists of a set of firms, I , each of which is a claim to an uncertain and perpetual dividend stream, c_i^t , where $i \in I$ and $t \geq \tau_i$. $\tau_i \in \mathcal{T}$ is the date of the initial cashflow of the firm. Thereafter the cashflows of the firms follow a lognormal random walk:

$$c_i^{t+1} = c_i^t \exp \left[\mu_i - \frac{1}{2} \sigma_i^2 + \sigma_i \epsilon_i(t+1) \right] \quad (1)$$

with $c_i^{\tau_i} = C_i$ and where the sequence $\{\epsilon_i(t), t \geq \tau_i\}$ is *iid* and $N(0, 1)$. For any two firms, i and j , let the correlation between $\epsilon_i(t)$ and $\epsilon_j(t)$ be denoted ρ_{ij} . Assume that C_i , τ_i , σ_i , μ_i and ρ_{ij} are all non-negative and have (cross-sectional) distributions that are independent of one another. That is, each realization represents a one time (non-negative) draw from the cross-sectional distribution of that characteristic that is independent of everything else in the model.⁴ For expositional clarity we will distinguish cross-sectional moments (as opposed to time-series moments) with a superscript ‘c’. For example, if x_i is an arbitrary characteristic of firm i , then the cross-sectional variance (across firms) of this characteristics is denoted $\text{var}^c(x_i)$.

The above assumptions are certainly plausible if not standard. Most models do not restrict the cross-section distribution of variables such as the initial cashflow, average cashflow growth and the standard deviation of cashflows. There is no obvious reason why the size of the initial cashflow should be related to the variance, covariance or subsequent growth of the cashflows. Nor should date of the first cashflow matter.

For any $t > \tau_i$, if $\sigma_i(t) \equiv \sigma_i \sqrt{t - \tau_i}$ and $\mu_i(t) \equiv \mu_i (t - \tau_i)$, then from (1) we get an expression for c_i^t as a function of C_i :

$$c_i^t = C_i \exp \left[\mu_i(t) - \frac{1}{2} \sigma_i^2(t) + \sigma_i(t) \psi_i(t) \right] \quad (2)$$

where $\psi_i(t) \equiv \frac{1}{\sqrt{t - \tau_i}} \sum_{\tau = \tau_i + 1}^t \epsilon_i(\tau)$ is $N(0, 1)$ and depends only on information between τ_i and t .

⁴More formally, if $f_x(\cdot)$ is the cross-sectional density function of firm characteristic x , then the joint density function of any two characteristics, say C_i and σ_i , is $f_{C_i}(\cdot) f_{\sigma_i}(\cdot)$.

The firms are traded on a spot markets at each time t . Since I am constructing an economy in which risk and return are unrelated I will assume that the dividend stream plays no part in determining the firm's price, p_i^t , on these spot markets. Instead, at each time t , prices are determined randomly, that is, each firm i 's price on every spot market is drawn independently of everything else in the model. The total return is then given by

$$R_i^t \equiv \frac{c_i^{t+1} + p_i^{t+1}}{p_i^t}.$$

If the mean of the price distribution is denoted p , then the expected return is:

$$E_t[R_i^t] \equiv \frac{\bar{c}_i^{t+1} + p}{p_i^t}.$$

This economy represents perhaps starkest alternative to the kind of economy in which the current asset pricing paradigms are applied. Not only are prices unrelated to risk, they are unrelated to every aspect of the firm. Green men on Mars might as well be pricing assets in this economy. Yet, if an econometrician were to calculate the beta of any stock with any arbitrary portfolio that requires no short sales then she would observe a positive relation between these betas and expected returns. To demonstrate why, take any portfolio in this economy that contains no short positions, denote it m , and consider pricing all other stocks off stock m . By definition,

$$R_m^t = \sum_{k \in I} \alpha_k R_k^t.$$

where the portfolio weights, α_k sum to 1 and are non-negative. In the single beta model, the beta (at time t) of stock i is defined to be

$$\begin{aligned} \beta_i^t &\equiv \frac{\text{cov}(R_m^t, R_i^t)}{\text{var}(R_m^t)} = \frac{\text{cov}(\sum_{k \in I} \alpha_k R_k^t, R_i^t)}{\text{var}(R_m^t)} \\ &= \sum_{k \in I} \alpha_k \frac{\text{cov}(R_k^t, R_i^t)}{\text{var}(R_m^t)} \\ &= \sum_{k \in I} \alpha_k \frac{\text{cov}(c_i^{t+1} + p_i^{t+1}, c_k^{t+1} + p_k^{t+1})}{\text{var}(R_m^t) p_i^t p_k^t} \\ &= \sum_{k \in I} \alpha_k \frac{\text{cov}(c_i^{t+1}, c_k^{t+1})}{\text{var}(R_m^t) p_i^t p_k^t} \end{aligned}$$

$$\begin{aligned}
&= \sum_{k \in I} \alpha_k \frac{c_i^t e^{\mu_i} c_k^t e^{\mu_k} \text{cov} \left(e^{-\frac{1}{2}\sigma_i^2 + \sigma_i \epsilon_i(t+1)}, e^{-\frac{1}{2}\sigma_k^2 + \sigma_k \epsilon_k(t+1)} \right)}{\text{var}(R_m^t) p_i^t p_k^t} \\
&= \sum_{k \in I} \alpha_k \frac{c_i^t e^{\mu_i} c_k^t e^{\mu_k} (e^{\sigma_i \sigma_k \rho_{ik}} - 1)}{\text{var}(R_m^t) p_i^t p_k^t} \\
&= \sum_{k \in I} K_k \frac{c_i^t e^{\mu_i} (e^{\sigma_i \sigma_k \rho_{ik}} - 1)}{p_i^t}, \tag{3}
\end{aligned}$$

where $K_k \equiv \frac{\alpha_k c_k^t e^{\mu_k}}{\text{var}(R_m^t) p_k^t} > 0$. Next, compute the *unconditional* cross-sectional correlation of the time- t beta and the time- t expected return:

$$\begin{aligned}
\text{cov}^c \left(E_t[R_i^t], \beta_i^t \right) &= \sum_{k \in I} K_k \text{cov}^c \left(\frac{\bar{c}_i^{t+1} + p}{p_i^t}, \frac{c_i^t e^{\mu_i} (e^{\sigma_i \sigma_k \rho_{ik}} - 1)}{p_i^t} \right) \\
&= \sum_{k \in I} K_k \text{cov}^c \left(\frac{c_i^t e^{\mu_i}}{p_i^t}, \frac{c_i^t e^{\mu_i} (e^{\sigma_i \sigma_k \rho_{ik}} - 1)}{p_i^t} \right) + K_k \text{cov}^c \left(\frac{p}{p_i^t}, \frac{c_i^t e^{\mu_i} (e^{\sigma_i \sigma_k \rho_{ik}} - 1)}{p_i^t} \right) \\
&= \sum_{k \in I} K_k \left(E^c \left[\left(\frac{c_i^t e^{\mu_i}}{p_i^t} \right)^2 (e^{\sigma_i \sigma_k \rho_{ik}} - 1) \right] - E^c \left[\frac{c_i^t e^{\mu_i}}{p_i^t} \right] E^c \left[\frac{c_i^t e^{\mu_i}}{p_i^t} (e^{\sigma_i \sigma_k \rho_{ik}} - 1) \right] \right) \\
&\quad + p K_k \left(E^c \left[\frac{c_i^t e^{\mu_i} (e^{\sigma_i \sigma_k \rho_{ik}} - 1)}{(p_i^t)^2} \right] - E^c \left[\frac{1}{p_i^t} \right] E^c \left[\frac{c_i^t e^{\mu_i} (e^{\sigma_i \sigma_k \rho_{ik}} - 1)}{p_i^t} \right] \right) \tag{4}
\end{aligned}$$

The terms in parenthesis in (4) are positive. To see why, note that by the independence of p_i^t ,

$$\begin{aligned}
&E^c \left[\frac{c_i^t e^{\mu_i} (e^{\sigma_i \sigma_k \rho_{ik}} - 1)}{(p_i^t)^2} \right] - E^c \left[\frac{1}{p_i^t} \right] E^c \left[\frac{c_i^t e^{\mu_i} (e^{\sigma_i \sigma_k \rho_{ik}} - 1)}{p_i^t} \right] \\
&= E^c \left[c_i^t e^{\mu_i} (e^{\sigma_i \sigma_k \rho_{ik}} - 1) \right] E^c \left[\frac{1}{p_i^t} \right]^2 - \left(E^c \left[\frac{1}{p_i^t} \right] \right)^2 E^c \left[c_i^t e^{\mu_i} (e^{\sigma_i \sigma_k \rho_{ik}} - 1) \right] \\
&= E^c \left[c_i^t e^{\mu_i} (e^{\sigma_i \sigma_k \rho_{ik}} - 1) \right] \text{var}^c \left(\frac{1}{p_i^t} \right) > 0, \tag{5}
\end{aligned}$$

where the inequality follows from the fact that c_i^t , μ_i and ρ_{ik} are all strictly positive. Next, note that

$$\begin{aligned}
&E^c \left[\left(\frac{c_i^t e^{\mu_i}}{p_i^t} \right)^2 (e^{\sigma_i \sigma_k \rho_{ik}} - 1) \right] \\
&= E^c \left[\left(\frac{C_i \exp \left[\mu_i(t+1) - \frac{1}{2}\sigma_i^2(t) + \sigma_i(t)\psi_i(t) \right]}{p_i^t} \right)^2 (e^{\sigma_i \sigma_k \rho_{ik}} - 1) \right]
\end{aligned}$$

$$\begin{aligned}
&= E^c \left[\frac{C_i e^{\mu_i(t+1)}}{p_i^t} \right]^2 E^c \left[E^c \left[\exp \left[-\sigma_i^2(t) + 2\sigma_i(t)\psi_i(t) \right] (e^{\sigma_i\sigma_k\rho_{ik}} - 1) \middle| \sigma_i, \tau_i, \rho_{ik} \right] \right] \\
&= E^c \left[\frac{C_i e^{\mu_i(t+1)}}{p_i^t} \right]^2 E^c \left[(e^{\sigma_i\sigma_k\rho_{ik}} - 1) E^c \left[\exp \left[-\sigma_i^2(t) + 2\sigma_i(t)\psi_i(t) \right] \middle| \sigma_i, \tau_i, \rho_{ik} \right] \right] \\
&= E^c \left[\frac{C_i e^{\mu_i(t+1)}}{p_i^t} \right]^2 E^c \left[e^{\sigma_i^2(t)} (e^{\sigma_i\sigma_k\rho_{ik}} - 1) \right]. \tag{6}
\end{aligned}$$

Similarly,

$$\begin{aligned}
&E^c \left[\frac{c_i^t e^{\mu_i}}{p_i^t} \right] E^c \left[\frac{c_i^t e^{\mu_i}}{p_i^t} (e^{\sigma_i\sigma_k\rho_{ik}} - 1) \right] \\
&= E^c \left[\frac{C_i \exp \left[\mu_i(t+1) - \frac{1}{2}\sigma_i^2(t) + \sigma_i(t)\psi_i(t) \right]}{p_i^t} (e^{\sigma_i\sigma_k\rho_{ik}} - 1) \right] E^c \left[\frac{C_i \exp \left[\mu_i(t+1) - \frac{1}{2}\sigma_i^2(t) + \sigma_i(t)\psi_i(t) \right]}{p_i^t} \right] \\
&= \left(E^c \left[\frac{C_i e^{\mu_i(t+1)}}{p_i^t} \right] \right)^2 E^c \left[\exp \left[-\frac{1}{2}\sigma_i^2(t) + \sigma_i(t)\psi_i(t) \right] (e^{\sigma_i\sigma_k\rho_{ik}} - 1) E^c \left[\exp \left[-\frac{1}{2}\sigma_i^2(t) + \sigma_i(t)\psi_i(t) \right] \right] \right] \\
&= \left(E^c \left[\frac{C_i e^{\mu_i(t+1)}}{p_i^t} \right] \right)^2 E^c \left[(e^{\sigma_i\sigma_k\rho_{ik}} - 1) E^c \left[\exp \left[-\frac{1}{2}\sigma_i^2(t) + \sigma_i(t)\psi_i(t) \right] \middle| \sigma_i, \tau_i, \rho_{ik} \right] \right] \\
&= \left(E^c \left[\frac{C_i e^{\mu_i(t+1)}}{p_i^t} \right] \right)^2 E^c \left[(e^{\sigma_i\sigma_k\rho_{ik}} - 1) \right]. \tag{7}
\end{aligned}$$

Using (6) and (7), the first term in paranthesis in (4) becomes,

$$\begin{aligned}
&E^c \left[\left(\frac{c_i^t e^{\mu_i}}{p_i^t} \right)^2 (e^{\sigma_i\sigma_k\rho_{ik}} - 1) \right] - E^c \left[\frac{c_i^t e^{\mu_i}}{p_i^t} \right] E^c \left[\frac{c_i^t e^{\mu_i}}{p_i^t} (e^{\sigma_i\sigma_k\rho_{ik}} - 1) \right] \\
&= E^c \left[\frac{C_i e^{\mu_i(t+1)}}{p_i^t} \right]^2 E^c \left[e^{\sigma_i^2(t)} (e^{\sigma_i\sigma_k\rho_{ik}} - 1) \right] - \left(E^c \left[\frac{C_i e^{\mu_i(t+1)}}{p_i^t} \right] \right)^2 E^c \left[(e^{\sigma_i\sigma_k\rho_{ik}} - 1) \right] \\
&> E^c \left[\frac{C_i e^{\mu_i(t+1)}}{p_i^t} \right]^2 E^c \left[(e^{\sigma_i\sigma_k\rho_{ik}} - 1) \right] - \left(E^c \left[\frac{C_i e^{\mu_i(t+1)}}{p_i^t} \right] \right)^2 E^c \left[(e^{\sigma_i\sigma_k\rho_{ik}} - 1) \right] \\
&= \text{var}^c \left(\frac{C_i e^{\mu_i(t+1)}}{p_i^t} \right) E^c \left[(e^{\sigma_i\sigma_k\rho_{ik}} - 1) \right] > 0. \tag{8}
\end{aligned}$$

Using (5) and (8) in (4) implies that for any stock i , $\text{cov}^c(E_t[R_i^t], \beta_i^t) > 0$.

In this economy in which risk has no role in stock prices, any portfolio that does not require short selling will appear to price stocks. Thus, the mere existence of an empirical factor model

is not sufficient to conclude that cross-sectional differences in expected returns are risk based. Furthermore, the portfolios themselves need provide no economic insight on the underlying reasons for the cross-sectional variance in security returns.

4 What do we Learn from the Additional Explanatory Power of Market Value?

One of the distinguishing features of both the Fama and French (1992) and Jegadeesh (1992) studies is that they find that the size related measures have additional explanatory power over and above the single beta model. At first glance this result might seem inconsistent with the single beta model and so might legitimately have a claim to the term “anomaly.” However, this result could be observed in an economy in which an asset pricing model such as the CAPM held perfectly.

Consider the one-period world in which all firms have the same expected cashflows. Assume that a researcher attempts to test an asset pricing model that in fact holds *exactly* in this world. Unfortunately for the researcher, assume that his empirical specification of the model is flawed, so although he has the right model, the empirical specification of the model does not precisely explain all risk factors. Thus the risk can be decomposed into two parts, the part explained by (the empirical specification of) the model, and the part that is unexplained by the model. However, since market value (book-to-market) is correlated with all risk factors, it will be correlated to any risk factor not explained by the model. Thus market value (book-to-market) will provide additional explanatory power in this test. Furthermore, so long as the asset pricing model itself does not predict a positive correlation between expected cashflow and expected return, I show elsewhere (Berk (1995)) that this result can be extended to any world in which expected cashflows are not positively correlated to expected returns.

Fama and French (1992) themselves concede that there is an error-in-variables bias in the empirical technique that they use. This being the case, it is therefore known, *a priori* that the empirical specification contains errors. Thus the additional explanatory power of market value should be expected and cannot be regarded as inconsistent with the single beta model.

These arguments imply market value (book-to-market) might be a useful addition to any study that uses an asset pricing model to predict stock returns. Since market value explains any unmeasured risk, it can be used as a measure of how much of the expected return remains unexplained by the model. If a portfolio manager claims to have an asset pricing specification that predicts

expected returns perfectly then, at a minimum, it must leave any market value related measure with no residual explanatory power. As such, the market value related variables loom as natural yardsticks by which all asset pricing models could potentially be measured.

A note of caution is, however, in order. It is important that researchers carefully derive the theoretical implication of using these variables in their tests. For example, one could construct a test of a specific asset pricing model by grouping stocks into subsets based on the risk premium as predicted by the model under consideration. If the model is correct and the empirical specification does not contain errors, then variations of actual returns within each subset should be unpredictable and unrelated to risk. This implies that market value (book-to-market) will have no explanatory power within each subset and so this condition could form the basis of an empirical testing procedure.

Now consider repeating the above procedure, except reversing the order. Rather than sorting stocks by the predicted risk premium, sort stocks into subsets by market value. Then test the asset pricing model within each subset. Even if the asset pricing model and the empirical specification contains no errors, it is quite possible that a researcher will find that the model has no explanatory power. This follows because, as we have argued, market value is already a measure of the risk premium so the initial sort already grouped stocks into subsets of similar risk premia! The cross-sectional variation in return within a subset cannot be risk related, and so is unpredictable by the asset pricing model. The above test could lead to a rejection of an asset pricing model that in fact held perfectly. Clearly, such a procedure is an inappropriate use of market value in an asset pricing test. Yet, Daniel and Titman (1995), for example,⁵ use precisely this procedure to conclude that no factor model can explain asset returns. This study provides a useful lesson for future researchers — if the tautological relation between market value (book-to-market) and expected return is not properly taken into account in the empirical specification, incorrect inferences can be drawn.

5 Conclusion

The object of this chapter was to provide a perspective on three important issues in financial economics. First I have argued that the empirically observed size effect should not be considered an anomaly. Second, while asset pricing model that are motivated purely by the empirically observed

⁵I choose this particular study to illustrate my point solely because it is a topical study that I am familiar with (having recently discussed it at the 1996 AFA meetings).

relation between market value (book-to-market) and return might well do a good job capturing cross-sectional variation in expected returns, they provide no insight on the *economic causes* of this cross-sectional variation. In particular, the variation need not result from cross-sectional variation in firm riskiness. Finally, while the additional explanatory power of market value (book-to-market) over asset pricing models such as the CAPM or APT clearly indicate that, at least as far these empirical specifications are concerned, these models cannot be capturing all cross-sectional variation in expected returns, this result cannot be used, by itself, to reject the model. It is quite conceivable that the result could derive from errors in the empirical specification rather than shortcomings in the model under consideration.

References

- Ball, R., (1978), "Anomalies in relationships between securities' yields and yield-surrogates," *Journal of Financial Economics*, **6**:103-126.
- Banz, R.F., (1981), "The relationship between return and market value of common stocks" *Journal of Financial Economics*, **9**:3-18.
- Basu, S., (1983), "The relationship between earnings yield, market value, and return for NYSE common stocks: Further evidence," *Journal of Financial Economics*, **12**:129-156.
- Berk, J.B. (1995) "A Critique of Size Related Anomalies," *Review of Financial Studies*, **8**:275-86.
- Berk, Jonathan B., Richard C. Green and Vasant Naik (1996), "Optimal Investment, Growth Options and Security Returns," working paper, available via <http://berk.commerce.ubc.ca>.
- Chan, L. K., Y., Hamao and J. Lakonishok (1991), "Fundamentals and stock returns in Japan," *Journal of Finance*, **46**:1739-1789.
- Chen, N., S.A. Ross and R. Roll (1983), "Economic Forces and the Stock Market," *Journal of Business*, **59**:383-403.
- Daniel, K. and S. Titman (1995) "Evidence on the Characteristics of Cross Sectional Variation in Stock Returns," Working Paper, University of Chicago.
- Dimson E. ed. *Stock Market Anomalies*, Cambridge University Press, Cambridge, U.K.
- Fama, E.F., and K.R. French (1992), "The Cross-Section of Expected Stock Returns," *Journal of Finance*, **47**:427-466
- Fama, E.F., and K.R. French (1993), "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, **33**:3-56
- Fama, E.F., and K.R. French (1995), "Size and Book-to-Market Factors in Earnings and Returns," *Journal of Finance*, **50**:131-84 .
- Ferson, W.E., (1995), "Warning: Attribute-sorted Portfolios Can be Hazardous to Your Research!," forthcoming in Saitou, S., K. Sawaki and K. Kubota, eds., *Modern Finance Theory and Applications*, Gakajyutsu Shuppan Center, Osaka, Japan.

- Jegadeesh, N (1992), "Does Market Risk Really Explain the Size Effect," *Journal of Financial and Quantitative Analysis*, **27**:337-352.
- Green, Richard C., (1986), "Benchmark Portfolio Inefficiency and Deviations from the Security Market Line," *Journal of Finance*, **41**: 1051-68.
- Keim, D.B., (1983), "Size-related anomalies and stock return seasonality: Further empirical evidence," *Journal of Financial Economics*, **12**:13-32.
- Keim, Donald B. and Robert F. Stambaugh (1984), "A further investigation of the weekend effect in stock returns," *Journal of Finance*, **39**:819-840.
- Kothari, S. P. and J. Shanken (1995), "Book-to-market, Dividend Yield, and Expected Market Returns: A Time-series Analysis," working paper, University of Rochester, June 1995.
- Kuhn, T. (1970), *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago, USA.
- Lee, C.M.C., A. Shleifer and R.H. Thaler (1991), "Investor Sentiment and the Closed-End Fund Puzzle," *Journal of Finance*, **46**:75-110.
- Pontiff, J. and L. Schall (1995), "Book-To-Market as a Predictor of Market Returns," working paper, University of Washington, Seattle, June 1995
- Reinganum, M.R., (1981), "Misspecification of capital asset pricing: Empirical anomalies based on earnings' yields and market values," *Journal of Financial Economics*, **9**:19-46.
- Rosenberg, B., K. Reid and R. Lanstein (1985), "Persuasive evidence of market inefficiency," *Journal of Portfolio Management*, **11**:9-17.
- Roll, R., (1977), "A Critique of the Asset Pricing Theories Tests; Part 1: On Past and Potential Testability
- Stattman, D., (1980), "Book values and stock returns," *The Chicago MBA: A Journal of Selected Papers*, **4**:25-45.
- Ziembra, W.T., 1994, "World Wide Security Market Regularities," *European Journal of Operational Research*, **74**, 198-229.