A Study into Certain Aspects of the Cost of Capital for Regulated Utilities in the U.K.

Stephen Wright
Birkbeck College and Smithers & Co

Robin Mason
University of Southampton and CEPR

David Miles
Imperial College and CEPR

On Behalf Of:
Smithers & Co Ltd
20 St Dunstan's Hill
London EC3R 8HY

February 13, 2003

This report was commissioned by the U.K. economic regulators and the Office of Fair Trading.

The U.K. economic regulators are The Civil Aviation Authority (CAA), Office of Water Services (OFWAT), Office of Gas and Electricity Markets (Ofgem), Office of Telecommunications (Oftel), Office of the Rail Regulator (ORR) and Office for the Regulation of Electricity and Gas (OFREG).

Any opinions contained in this report are those of the authors, not of the economic regulators or the OFT.

Contents

List of Figures

List of Tables

Acknowledgements

We would like to thank Stephen Satchell for very useful comments on the first draft; he is not responsible for any errors that this report may contain. We would also like to thank Spyridon Andreopoulos, Yen-Ting Hu and Eleonora Patacchini for excellent research assistance.

1. Introduction and Executive Summary

1.1. Background to the Study

The cost of capital is a key input in the regulatory process for utilities. The utility regulators are responsible for regulating a wide range of industries—gas, electricity, water, telecom, rail, airports and postal services. An important function of each of the regulators is to set price limits for those parts of those industries where firms have significant monopoly power in setting prices faced by customers. In setting these price limits, regulators need to decide what would constitute a "fair" rate of profit.

To do this regulators need to assess the return that investors in these firms would have earned if they invested in any firm with a comparable level of risk. In contrast to their product markets, all utilities can reasonably be assumed to face fully competitive capital markets. In such markets, asset prices will always ensure that a new investor in the firm will simply earn the competitive (risk-adjusted) return. If the rate of profit earned by the utility exceeds that on competitive firms, this will simply be reflected in a higher market valuation of the firm. This will only be to the benefit of existing shareholders to the extent that the excess profits could not have been rationally forecast at the time they invested in the company.

Thus, the limited nature of competition in product markets is not a relevant issue in assessing the cost of capital. In capital markets, the firm can be treated as a "price-taker" rather than a "price-maker". All that matters is the market cost of capital adjusted for the firm's risk characteristics, irrespective of whether it is a monopoly.

There is and has been much research and debate concerning the cost of capital for businesses in general, let alone simply for utilities. There have been a number of new develop-

ments in the finance literature, while at the same time, regulators have had to address new and complex issues when estimating the cost of capital. It is frequently difficult for regulators to address these issues in any depth on an ad hoc basis. The regulators have also attracted criticism, whether justified or not, for not appearing to be consistent on their approach to the cost of capital. This criticism has come from the industries themselves, various ad hoc groups (e.g., The Better Regulation Task Force) and certain government departments.

The regulators have decided, therefore, to commission this research project into certain aspects of the cost of capital in order to gain an independent view on emerging and new issues in the estimation of the cost of capital, the scope for greater consistency between regulators and to understand why there may be differences in approach.

1.2. Overview of this Study

Our study is not intended to provide exhaustive coverage of every possible aspect of the cost of capital. Instead, at the request of the regulators, we have focussed on certain key areas, as follows:

Chapter 2 examines the "common components" of the cost of equity capital: i.e., those that are common to all firms, and all competing asset pricing models.

Chapter 3 provides a comparison of asset pricing models for regulation: we re-examine the standard "CAPM" approach, in the light of recent academic work on alternative asset pricing models, with a focus on the relevance for practical implementation in regulation.

Chapter 4 focusses on practical issues in estimation of asset pricing parameters for utilities, with particular focus on estimation of the CAPM "beta".

Chapter 5 discusses the case for consistency in setting the cost of capital: whether, even if regulators share the same central approach to estimating the cost of capital, uncertainty as to the true value may lead regulators to set different values in different industries.

Chapter 6 discusses "regulatory risk": whether there may be risks associated with regulated industries that are not captured by standard measures of systemic risk, but the impact

of which on investment returns cannot be eliminated by diversification.

1.3. Executive Summary

Since this report is quite lengthy, and at points fairly technical, we provide below a short summary of our key conclusions. In addition, each chapter also ends with a list of the key conclusions relevant to that particular chapter.

1.3.1. The Common Components of the Cost of Equity

Our starting point in Chapter 2, as in all such studies, is the "weighted average cost of capital": the rate of return required by a company's investors, whether in its debt or its equity. Our primary focus in this study, however, is on issues relating to the cost of equity capital, in relation to which there is the greatest degree of difficulty in measurement, and (largely as a result) the greatest degree of controversy.

All firms are different. But a central element in finance is that, despite their differences, a significant element in the cost of raising equity finance is common to all firms. In the most commonly used framework for asset pricing, the Capital Asset Pricing Model (CAPM) of Sharpe (1964) and Lintner (1965), there are just two determinants of the expected return on any asset: the return on a riskless asset; and the expected gap between the "market return" and the risk-free rate. In the CAPM the only thing that is specific to any given asset is its "beta" (β), that determines how responsive the asset's return is to the excess return on the market. While the CAPM itself has come in for considerable criticism in recent years(that we discuss at some length in Chapter 3) it is fair to say that virtually any asset pricing model that is used to derive an appropriate estimate of the cost of capital for regulated industries must inevitably build on estimates of the common components that feed into the CAPM itself. Our starting point, therefore, needs to be to examine how to estimate these common components.

It is standard usage to follow the CAPM by building up the cost of equity capital from its two common elements: the risk-free rate and the expected excess return on the market (or

equity premium). We argue however that this approach is not necessarily the most efficient way to proceed. In the CAPM, the expected return on a firm's equity can be re-expressed equivalently as a weighted average of the risk-free rate and the expected market return, where the closer is a given firm's β to unity (i.e., the closer it is to being "average") the lower the implied weight on the safe rate. Regulated industries are unlikely to be "precisely" average, with a beta of unity; but nonetheless the dominant element in their cost of capital will always be the expected market return, with a distinctly smaller role for the risk-free rate. This will also generally be the case even in alternative, more complicated asset pricing models.

The relatively greater importance of the market return is fortunate for the regulators, since we argue that there is considerably more uncertainty about the true historic risk-free rate, and hence the equity premium, than there is about the market return itself. The historic size of the equity premium is still the subject of considerable puzzlement and controversy amonst academics; but this is largely due to the historic behaviour of the risk-free rate (proxied by the short-term interest rate). In contrast, we summarise a range of evidence that the equity return has, over reasonably long samples, been fairly stable both over time, and across different markets.

Before examining the data, we note that care should be applied as to whether returns are being measured using arithmetic or "geometric" averaging. The former is conceptually superior, though possibly less stable. The most crucial thing is to be aware that the difference between the two measures can be significant—as much as two percentage points or more.

Both on *a priori* grounds, and on the basis of evidence, our strong view is that estimates of both the equity return and the risk-free rate should be formed on the basis of international evidence, not just from the UK experience. We examine a range of empirical issues and estimates. Our central estimate of the cost of equity capital, derived from a wide range of markets, is around 5.5% (geometric average), and thus 6.5% to 7.5% (arithmetic average). We cannot, however, be at all confident that these estimates are precisely correct: 95% confidence intervals are, at a conservative estimate, of up to two percentage points either side of the point estimates.

Problems in assessing historic mean values of the safe rate imply that estimates of the

future risk-free rate (that are, as we have noted, fortunately of distinctly lower importance for regulators) should probably be derived in a forward-looking way from current rates. However, in so doing, account should be taken of forecast future movements of short-term rates, derived both from market data and published forecasts. A common estimate of the equilibrium risk-free rate would be of the order of 2 1/2%. Using this figure, the implied equity risk premium is of the order of 3 percentage points (geometric) and 4–5 percentage points (arithmetic).

1.3.2. Asset Pricing Models

The Capital Asset Pricing Model (CAPM) is still widely used to estimate firms' costs of capital. There is considerable evidence of empirical shortcomings in the CAPM; but its clear theoretical foundations and simplicity contribute to its continuing popularity. In Chapter 3 we examine a range of competing models.

The CAPM is a linear model: i.e., the expected excess return on any given asset vs the risk-free rate is in fixed proportion to the expected excess return of the market, with the asset's "beta" determining the degree of proportionality. The alternative of nonlinear models of asset pricing (in which this proportionality does not hold) has not achieved the popularity of the CAPM. There are several reasons for this. The most important is the problem of 'data overfitting'—fitting the sample data "too well", so that both systematic and entirely random factors are explained by the model. Nonlinear models are particularly prone to this temptation. The problem is compounded by the absence of any one method that can test for the problem of overfitting. In addition, in many cases a nonlinear model can be approximated well by a linear model. Finally, recent research suggests that a carefully specified "conditional CAPM" —i.e., one in which the parameters of the model vary over time—will usually perform better than a nonlinear model.

Such conditional models, in which the parameters are time-varying, have been the focus of much recent work. As with nonlinear models, the problem of data overfitting is present; and there is no test to assess the extent of the problem. In addition, there is no test available to assess whether the way parameters are allowed to vary has been done correctly. Despite the large amount of work in the area, the methodology is some way from being agreed and

testable.

Probably the most popular competitor to the CAPM has been in the form of linear multifactor models. These, in contrast to the CAPM, suppose that there is more than one factor driving asset returns. Models along these lines have received considerable attention, particularly since the influential work of Fama and French. The standard difficulty with multifactor models is the satisfactory identification of the factors. As in the case of other competitors to the CAPM, multifactor models have been criticised for overfitting and data mining.There has been, for example, a considerable debate about whether the two additional factors that Fama and French have focussed on - one, an indicator of how returns vary with firm size, the other, of how returns vary with "value" (proxied by the ratio of book value to market value) - are robust in other time periods and markets. While both can be shown to have strong explanatory power for asset returns on a period by period basis, the key issue is whether these factors are, on average "priced": i.e., whether the risk premia associated with these factors are clearly statistically significant. We present some evidence that they are not, especially when the sample period is extended to include later data. If the additional factors do not have statistically significant risk premia, multfactor models reduce, in effect, to the CAPM.

In summary: the empirical shortcomings of the CAPM are known. Alternative models to address this issue have their own shortcomings—weak theoretical foundations and empirical challenges. In our view, there is at present no one clear successor to the CAPM for practical cost of capital estimation. We do however feel that alternative models provide helpful insights into the points of vulnerability of the CAPM, and may also provide information on the robustness of the CAPM beta.

1.3.3. Beta Estimation

In Chapter 4 we examine practical issues associated with estimating the CAPM "beta". We illustrate these issues with a practical example: we estimate β for British Telecom, using a range of different techniques.

There is a case to be made for using daily, or perhaps weekly, data rather than monthly

data in estimating beta. For a share where trading is not significantly thinner or thicker than for the market as a whole, using daily data has real advantages.

But where there may be a lag between the impact some events have on a particular share and the market in general going to lower frequency data can help. If one had to use the same estimation frequency for a very large number of different companies there is an argument that it makes sense to go to weekly or monthly data because some stocks really take time to catch up with general market news.

But regulators do not need to use the same frequency of data for estimates of different companies (unlike a commercial provider like the LBS which has standardised procedures and runs an automated service where all companies betas are calculated in the same way using 50 monthly observations). We conclude that using daily data may be right for many—but not all—companies

Adjusting standard errors for heteroskedasticity and serial correlation is important. Fortunately this is now a standard option in most econometric packages.

A case can be made that a portfolio which reflects the mix of assets of the typical stock holder in the company should be used as the "market portfolio". For large UK companies whose shares are largely held by UK investors this implies a market portfolio with about 70% of its weight on UK assets and 30% on overseas assets. All returns should be in sterling.

We also discuss the issue of "Bayesian adjustments" to beta, that allow for the fact that, in the absence of better information, the best guess for any firm's beta must be unity (the beta of an "average" firm). While such adjustments are correct in principle, in practice this may not make much difference if daily data are used because the resulting estimates of beta are typically very precise. With monthly data the Bayesian adjustment is likely to be more significant.

1.3.4. Consistency

Once the key elements of an asset pricing model (e.g., within the CAPM, the "common components" of the risk-free rate and the market return, together with the asset-specific

beta) have been estimated, they can be used to determine the appropriate form and level of regulation for a particular industry. We feel it is appropriate that regulators should ideally take a consistent approach to the way these key elements are measured.

Typically, however all elements are estimated with error: that is, any point estimate is accompanied by a range of uncertainty (or "confidence interval") which can be large. In Chapter 5 we ask: should the regulator use just the point estimate to set e.g., a price cap? Or should the uncertainty associated with the estimate also be reflected in the regulatory decision? And should all regulators respond to this uncertainty in the same way?

There are two extremes that the regulator will try to avoid. The first is setting the price cap too high, and so allowing the regulated firm to over-price. The second is setting the price cap too low, and so discouraging the regulated firm from undertaking efficient levels of investment, for example. If the first factor is the more important consideration for the regulator, then this should imply setting a low price cap. If the second factor is dominant, then the regulator will set a relatively high price cap. Which factor is the more important depends on the fine detail—the exact structure of costs and demand—of the industry.

These facts can be phrased in an alternative way by defining the *effective cost of capital estimate* to be the level of the cost of capital that should be used by the regulator in setting the price cap. A higher price cap corresponds to a higher effective cost of capital. Our analysis shows that the effective cost of capital estimate that should be used by a regulator will depend on demand and cost conditions, as well as the point estimate and error in cost of capital estimation. Therefore two regulators who share the same point estimate and confidence interval for the costs of capital for their regulated firms will, in general, choose different effective costs of capital for price cap purposes, to reflect the demand and cost characteristics of the firm that they regulate.

1.3.5. Regulatory Risk

In Chapter 6, we analyse three questions. First, what is the effect, if any, of regulatory inconsistency on a firm's cost of capital? Secondly, in what ways can different forms of regulation affect a firm's cost of capital? Finally, how will a firm react to regulation that

affects its cost of capital?

A common concern among those involved in regulation is that the regulator can itself introduce risk through unpredictable or unjustifiable regulatory intervention. We argue that this concern is largely misplaced. The central message of asset pricing theory is that only factors that co-vary with some systematic risk factor (eg, the market portfolio in the CAPM) affect a firm's cost of capital. Hence 'regulatory risk' arises only when the regulator's actions introduce systematic (i.e., non-diversifiable) risk. Any regulatory action that has an effect that can be diversified does not contribute to risk. True regulatory risk arises only when the regulator takes actions that cause the returns of the firm to be correlated with some systematic risk factor. A prime example of this is when a regulator decreases a price cap in response to a macro-economic shock that increases the profit of a firm.

Nevertheless, regulation does affect a firm's cost of capital by affecting the way a firm's profits vary with undiversifiable risks (such as macro-economic shocks). We show, in a simplified model, that the beta of a firm that is subject to cost uncertainty is increased by price cap regulation. In contrast, the beta of a firm that is subject to demand uncertainty is decreased by price cap regulation. In both cases, the change in beta is a result of the fact that the firm's ability to respond to undiversifiable shocks is limited by the price cap regulation. Cost pass-through can mitigate the effect in the case of cost uncertainty. The firm itself can offset this effect of regulation through its choice of activities or projects. In fact, we show that a regulated firm when it is faced with cost uncertainty will tend to choose projects with lower betas; and, when faced with demand uncertainty, will tend to choose projects with higher betas.

One implication of this is the a firm's short-run and long-run betas are likely to be different. With cost uncertainty, a firm's short-run beta (when it is less able to choose its project freely) will be relatively high; over the long-run, however, when it is able to choose its project, its beta will be lower. (The converse holds for demand uncertainty.)

An important issue highlighted by this analysis is that the type of uncertainty faced by a firm is an important determinant of the effect of regulation on its beta. Only non-diversifiable risk is important; and it is crucial to know whether the risk is on mainly on the cost or demand side.

2. The Common Components of the Cost of Equity

2.1. The Weighted Average Cost of Capital

The weighted average cost of capital ($WACC_i$) for a given firm i, is the average rate of return required by a company's investors, whether in its debt or its equity. A standard way to express this is:

$$WACC_i = g_i.R_i^D + (1 - g_i)R_i^E \tag{2.1}$$

where: g_i is the proportion of debt finance (or 'gearing'); R_i^D is the required rate of return on debt;$1 - g_i$ is the proportion of equity finance; and R_i^E is the required rate of return on equity.

Both R_i^D and R_i^E are assumed to be required rates of return after tax (thus the required return on debt is net of any impact of tax shelter).

2.1.1. The Cost of Debt

This study does not focus to any great extent on the cost of debt, R_i^D, given the relatively minor problems of measurement, since most large quoted firms with non-zero gearing have sufficiently large volumes of marketable debt outstanding to enable at least a ballbark estimate of the firm-specific cost of debt. A number of caveats are however worth noting:

- Strictly speaking R_i^D should measure the (unobservable) expected return on the firm's

debt, which is not identical to the observable yield, due to default risk: ie

$$R_i^D = (1 - \pi)Y_i^D \tag{2.2}$$

where π is the probability of default, and Y_i^D is the observable yield on firm i's debt.. However, default probabilities can be estimated fairly straightforwardly from published default rates on debt of a similar credit rating: such data are readily available from credit rating agencies. For most regulated utilities this adjustment is likely to be fairly minor.

- Even after this adjustment, this measures the *average* cost of debt, when what is strictly required is the marginal cost. Theory would suggest that the marginal cost will be greater than average, due to increasing default risk as leverage rises. But, again, this form of bias is likely to be fairly small at the levels of gearing observed for most regulated firms.

- If assumptions are to be made over a reasonably long horizon, it is not appropriate simply to assume that the cost of debt will remain constant, since the general level of interest rates at the relevant maturity may well be forecast to move up or down in the future (for a further discussion of this issue, see Section 2.5.2 below). A further adjustment may therefore be required for future forecast movements in riskless rates (whether long- or short-term), holding the relevant premium constant.

- While the tax shelter on debt is normally calculated as simply 1− the corporation tax rate, it has been argued that this may overstate the tax bias compared to equities, since there may also be explicit or implicit subsidies to equity returns.[1]

2.1.2. The Cost of Equity

All firms are different. But a central element in finance is that, despite their differences, a significant element in the cost of raising equity finance is common to all firms. By way of illustration, the commonly used Capital Asset Pricing Model (CAPM) of Sharpe (1964) and

[1]This issue is however well dealt with in standard finance textbooks. See, for example, the treatment in Copeland and Weston (1992), Chapter 13.

Lintner (1965) (on which more below in Chapter 3) assumes that the cost of equity for firm i, is the expected return on investing in a single share in that firm, in turn given by:

$$E(R_i^E) = R_f + \beta_i \left(E(R_m) - R_f\right) \tag{2.3}$$

where R_f is the (assumed fixed) return on a safe investment, and R_m is the return on investing in a market index.

In the CAPM, therefore, the only element in the cost of equity specific to the firm is its "CAPM beta", β_i; that captures the sensitivity of the firm's equity to the "systematic" risk captured by the excess return on the market index. While any given firm will also have an element of idiosyncratic risk (i.e., not correlated with the market return), this will not, in equilibrium, be priced by the market.[2]

While the CAPM itself has come in for criticism (that we discuss at some length in Chapter 3) for overly simplifying the nature of responses to systematic risk, it is fair to say that virtually any asset pricing model that is used to derive an appropriate estimate of the cost of capital for regulated industries must inevitably build on estimates of the common components that feed into the CAPM itself.

2.1.3. Defining the Common Components of the Cost of Equity Capital

It is standard usage to follow the CAPM specification, as in (2.3) by building up the cost of equity capital from the two elements therein: the risk-free rate, R_f and the expected equity premium, $E(R_m - R_f.)$. A point that we shall stress at various points in this chapter is that this approach is not necessarily the most efficient way to proceed.

These two elements can, instead, identically be expressed in terms of of the two underlying

[2]The intuitive rationale usually presented for this is that, in a sufficiently well-diversified portfolio (i.e., in which portfolio shares become sufficiently small) , the impact of idiosyncratic risk on any individual investor is negligible. While this rationale is correct for some asset pricing models, such as the Arbitrage Pricing Theory of Ross (1976), it is not actually required in the CAPM, which derives its key features from an assumption of marginal pricing in the neighbourhood of a general market equilibrium in which all assets are willingly held. This requires the representative investor to have portfolio shares equal to asset shares in total market value – these need not necessarily be negligibly small.

returns, R_f and $E(R_m)$.The latter is of cost the cost of equity capital for the "average" firm (*i.e.*, a firm with a beta of unity in the CAPM).We shall argue that this alternative decomposition while, obviously definitionally identical, both provides important insights, and a more practical approach to calculating the cost of equity capital, for two key reasons.

First, in the the CAPM pricing equation, (2.3) the expected return on a firm's equity can be re-expressed equivalently as a weighted average of the the risk-free rate (with weight $1 - \beta_i$) and the expected market return (with weight β_i)

$$E(R_i^E) = (1 - \beta_i)R_f + \beta_i E(R_m) \tag{2.4}$$

thus, the closer is a given firm's β to unity (*ie*, the closer it is to being "average") the lower the implied weight on the safe rate. Regulated industries are unlikely to be "precisely" average, with a beta of unity; but nonetheless the dominant element in their cost of capital will always be the expected stock market return, with a distinctly smaller role for the risk-free rate. This will also generally be the case even in alternative, more complicated asset pricing models.[3].

Second, we shall argue that there is reason to view the expected market return (hence the average cost of equity) as both more explicable in terms of underlying theory, and more stable over long historical samples, than the return on "safe" assets. Given the relative weightings on the two returns implied by (2.4), this is, for our purposes, fortunate. But it also implies, as we shall see, that the standard practice of building up the average cost of equity by adding an estimate of the equity premium to an estimate of the safe rate may be, at best, a not particularly efficient way to proceed, and at worst, a source of misunderstanding and errors. We shall return to this theme later in the chapter.

[3]Indeed, as we discuss below, in Chapter 3, Fama & French's (1992,1996) empirical implementation of Ross's (1976) APT model tends to point to estimates of β_i for almost *all* firms that are typically insignificantly different from unity, implying no differential role at all for the risk-free rate.

As Cochrane (1997) has pointed out, neither the standard CAPM, nor its more recently developed alternative asset pricing models, are designed to *explain* the common components: these are simply used as inputs to such models. To find any such candidate explanation, it is necessary to look deeper at the fundamental determinants of asset prices: the "Consumption CAPM", or some variant thereof.

Since Mehra & Prescott (1985) and Weil (1989) it has been established that simple versions of the consumption CAPM model signally fail to explain observed values of either the risk-free rate or the equity premium. As we shall see below, however, a feature of the model that it is less frequently acknowledged is that there is no such clear-cut failure of the model to explain the market return (the cost of equity) itself.

In order to understand the nature of the "equity premium puzzle" and "risk-free rate puzzle" and the associated debate on their empirical magnitudes, it is worth briefly summarising the underlying basis in theory.

Mehra and Prescott's original paper in 1985 setting out the equity premium puzzle arose out of what seems, in retrospect, a very simple idea. They asked, in effect, whether historic returns on stocks and safe assets could be reconciled with a model of a single "representative investor" (whose consumption could be represented by the average consumption level in the whole economy) who maximises utility over time, by choosing between various alternative assets.

A common misperception of the relationship across different assets is that rates of return depend simply on the relative volatility of their rates of return. In fact, standard theory suggests that they should depend on the *correlation* of these returns with some benchmark. In the standard CAPM, this is the market return; in the consumption CAPM it is the marginal utility of consumption.

As in most of economics, the rationale for this comes from the idea that someone who is maximising their utility will do so by equalising, at the margin, the gains and losses from any small change in anything they choose. If the gain from a small change would precisely

offset the loss, then, by implication, they must have the optimal amount of that particular thing.[4]

This general principle can be applied to the choice between consuming a pound today, and investing that same pound in any given asset for, say, a single year. The cost to giving up a pound today is straightforward, and known: it will be equal to the marginal utility of an extra pound's worth of consumption today. In an uncertain world, the gain from investing in the asset will be uncertain. It will be equal to the expected value of:

$$(1+R_j) \times \text{ marginal utility of an extra pound of consumption next year}$$

where R_j is the return on the jth asset next year.

If all investors can be represented by a single representative investor, then the expected return on asset j set by financial markets must be just high enough to make that investor indifferent between holding the asset and not holding it.

If there were no uncertainty, then all assets would offer the same return, that would simply be determined by the ratio of the marginal utility of an extra unit of consumption today to the marginal utility of an extra unit of consumption next year.

With uncertainty, things are more complicated. If the return on asset j is uncertain, the return a representative investor would have been happy with in a certain world may not be enough. This will be the case if, for example, the return on the asset and the consumer's marginal utility next year are negatively correlated. This will depress the expected benefit from investing in that asset, since times when the rate of return is high will be times when the marginal utility of an extra unit of consumption is low, and vice versa. Since it is standard to assume that marginal utility declines as consumption rises, the counterpart to this is that assets that yield higher-than-expected returns when consumption is higher than expected (hence when an extra unit of income is less valuable) must offer higher returns on average.

The intuition for this is that such assets are the opposite of an insurance policy for consumption: they pay out when consumption is already high, rather than when it is low.

[4] Assuming that preferences are "well-behaved", i.e. that there is a unique optimal choice.

Risk-averse individuals will normally pay for insurance; they therefore should rationally expect to be rewarded for "anti-insurance".

This offers a rationale for why stocks should yield higher returns than safe assets (which, if they are truly safe, should have a zero correlation with consumption) since good news for consumption is normally accompanied by good news for the economy as a whole, and hence usually implies good returns on shares.

Mehra and Prescott's insight was to point out that while available data were consistent with the qualitative features predicted by theory, the model they used did not offer a rationale for the *magnitude* of the gap between returns on stocks and safe assets. The observed covariance between stock returns and aggregate consumption should essentially be the only determinant of the equity premium. But this covariance is rather low: hence the only way that the equity premium can be explained is by assuming that marginal utility varies a lot in response to small changes in consumption. By implication, risk aversion must be very high.

A simple measure of risk aversion, used by Mehra and Prescott and most subsequent researchers, is the coefficient of relative risk aversion (often called γ). This captures the extent to which the representative consumer wants to have smooth consumption, both over time, and in different "states of nature". Someone who wants very smooth consumption will be very risk-averse – they will be prepared to pay a lot to insure themselves against fluctuations in their consumption. The higher is γ, the less the consumer values any marginal increase in consumption above this stable level., and hence the more they will pay for insurance. They will also be less willing to substitute consumption across different points in their life.[5] Most economists have concluded that a reasonable value for γ is of the order of 1 to 3 or 4, based on observed attitudes to risk in other contexts. But to match precisely the observed equity premium in Mehra & Prescott's sample period, γ would need to have been of the order of 18 or more.[6]

While significant (and reasonable) doubt has been cast on the original estimates of the equity premium used by Mehra and Prescott (now generally agreed to be an over-estimate),

[5] In the standard model, the elasticity of intertemporal substitution is given by $1/\gamma$, so a higher value of γ implies a preference for smoother consumption over time.

[6] For helpful numerical illustrations of both puzzles, see Kocherlakota (*op cit*) or Campbell, Lo and MacKinlay (1997, Chapter 8).

Kocherlakota (1996), Campbell (2001a) and many others have concluded that the "equity premium puzzle" still remains, since all historically based estimates of the premium are significantly higher than the fraction of a percentage point that would be predicted in the Mehra-Prescott model.

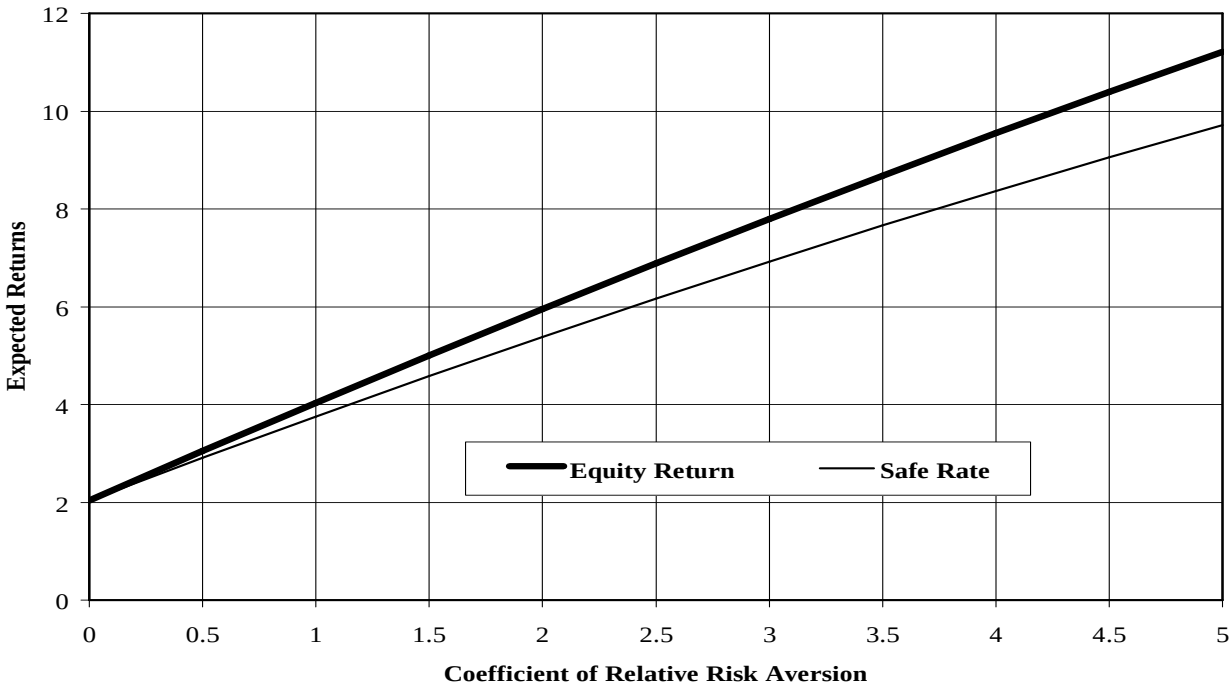
Alongside the equity premium puzzle, Weil (1989) established that there is an equivalent degree of puzzle about the determination of the risk-free rate, since in the standard model, the high required values of γ can only be reconciled with rather low historic average risk-free rates if the representative investor is assumed not to have a relative preference for consumption today over consumption tomorrow, as in the standard model, but to have the reverse order of preference – an assumption that appears massively counter-intuitive. This is the "risk-free rate puzzle".

It can indeed be argued (as, for example, does Kocherlakota, 1999) that the equity premium puzzle and the risk-free rate puzzle are very close to being two sides of the same coin. While the standard theory applied by Mehra and Prescott has major problems explaining the *relative* historic returns on equities vs safe investments, this is largely because it fails to explain the low absolute returns on safe assets. In contrast, it is not particularly hard to derive estimates of the expected stock return itself that are consistent with the theory.

Figure 2.1 illustrates, using data on US consumption growth, the volatility of stock returns, and the covariance of aggregate consumption with stock returns, taken from Campbell, Lo and MacKinlay (1997), and their log-linear calibration of the Mehra-Prescott model. The chart plots the predicted expected returns on stocks and the safe asset, for different values of γ, the coefficient of relative risk aversion.[7] The chart shows that it is quite easy to derive expected stock returns consistent with historical averages (discussed further below), for fairly modest values of γ. But it also illustrates the joint nature of the equity premium and risk-free rate puzzles: the implied risk-free rate is much higher, and hence the implied premium is much lower, than in the data.

The intuition behind the chart is that, supposing investors were risk-neutral (γ=0, or linear utility), but still had a relative preference for consumption today over consumption

[7] Using Campbell *et al*'s equations 8.25, 8.26, the assumption of a subjective discount rate (δ in Campbell *et al*) of 0.98, and the properties of log-normal distributions.



Assumptions: Subjective Discount Rate = 0.98; Sample Moments for US Stock Returns and Consumption data taken from Campbell et al (1997) Table 8.1

Figure 2.1: Expected Returns Implied by the Consumption CAPM

tomorrow (i.e., they were "impatient") This would imply that both stocks and the safe asset would still need to yield a positive return in equilibrium.

If risk aversion (i.e., concave per-period utility) enters the picture, it has two effects. The first is that, with growing consumption, investors, who, other things being equal, prefer a smooth path of consumption over time, need to be persuaded to postpone future consumption by higher returns. The more concave their utility, the more they need to be persuaded. This factor explains the upward slope of both lines. The second is that, at the same time, as noted above, the covariance of stock returns with consumption implies investors will need a *relative* rise in the stock return. While the chart shows this qualitative effect, the implied gap is not consistent with the data: the implied risk-free rate is much higher than in the data. Thus the chart illustrates both the equity premium and risk-free rate puzzles.

We should stress that this chart does not imply that the Consumption CAPM can nec-

essarily *explain* historic average stock returns: it simply says that there is no obvious inconsistency between data and theory. The clear inconsistency between theory and data relates to the safe rate, and hence the equity premium.

2.3. The Common Components: From Theory to Data and Back Again

The failure of the consumption CAPM to explain the equity risk premium and the risk-free rate, is at the heart of the empirical debate about their true values, since priors on true values play a key role in informing statistical inference. It is extremely rare that economic data will precisely match the predictions of theory; but economists tend to feel most comfortable when they can at least conclude that the theory is not clearly rejected. Thus theory may point to a given parameter, or combination of parameters having a particular value - zero or one, for example. Empirical estimates will never match these expectations precisely. But if, on standard tests, the hypothesised value of the "true" parameter cannot be rejected by the data, economists will usually act on the assumption that the best guess for the true parameter is its hypothesised value based on theory, rather than the actual historical estimate.

Unfortunately, in the case of the equity premium and the risk-free rate the only clear-cut prior in town has been so convincingly rejected, at least by historic data, that most economists[8] have concluded that this option is not open to them. To quote Dimson, Marsh, and Staunton (2001c)

> *Though some writers may give another impression, there is no single figure for the risk premium that theory says is 'correct'.*

The original Mehra & Prescott paper has spawned a vast academic literature, both theoretical and empirical, that has sought to resolve the puzzle either by casting doubt on the data, or by adjusting the model to be consistent with the data.

We shall discuss below a range of possible adjustments to data for the risk-free rate and the equity premium; but none of these are anywhere near large enough to resolve the original

[8]But, as we shall see below, not all.

puzzle, at least on the assumption that underlying behavioural parameters are reasonably stable (Kocherlakota, *op cit*; Campbell (2001a)).

The alternative approach, of adjusting the model to be consistent with the data, has spawned a number of papers that modify the Mehra & Prescott model along one dimension or another, and thereby attempt to "resolve" the puzzle. This was, indeed, what the original authors themselves had initially regarded as most likely:

> *Our conclusion is that most likely some equilibrium model with a friction will be the one that successfully accounts for the large average equity premium.*

Kocherlakota (*op cit*) and Campbell (2001a) provide useful summaries of the literature that has attempted to rise to this challenge. Examples are:

- Campbell (1993) examines the assumption that aggregate consumption is a poor proxy for the consumption of the true "representative investor" whose preferences are used to price financial assets. If the true representative investor is modelled as a rentier whose consumption path tracks changes in the value of the stock market precisely, this can in principle raise the covariance of consumption and market returns, and thereby lowers the required assumed value of γ to more plausible levels, in line with the earlier analysis of Friend and Blume (1975). [9]
- Campbell and Cochrane (1999) examine the impact of assuming that consumers form consumption "habits" that tend to make them far more risk averse at points when consumption is close to what they have come to regard as a minimum level of consumption: it thus generally implies a counter-cyclical equity risk premium. Unfortunately, this model singularly failed to explain the rise in the stock market in the 1990s.
- Constantinides, Donaldson, and Mehra (2002) show that high premia can result from liquidity constraints. Consumers who are early in the life cycle would ideally wish to borrow, but cannot do so: as a result, the safe rate is depressed, and the equity premium raised, below what it would be in the absence of credit constraints.

[9] However, this result only holds if market returns are unpredictable; allowing for predictability in returns (an issue discussed further below, in Section 2.4.5) pulls implied values of γ up significantly.

- Kocherlakota (*op cit*) discusses theories of the precautionary motive that may also help to explain the risk-free rate puzzle by introducing an additional motive for holding safe assets, arising from non-diversifiable labour income risk.

Although all of these models have some attractive and plausible features, none individually appears to provide a fully consistent resolution of the puzzles, leading Kocherlakota (*op cit*) to conclude that both remain puzzling.

A major problem associated with all these proposed resolutions is that, since the models used are, effectively, designed to match the data, they are very hard to test. The problem is not dissimilar to the "data mining" problem in other applied work. Economists know there is a puzzle in the data, and therefore seek to design models that explain away the puzzle: i.e., they effectively engage in "theory mining". If they did not know about the puzzle, they might look in other directions for modifications to their models, that might quite possibly imply that the puzzle would get bigger, not smaller.

It must be acknowledged that, in the face of such problems of inference, there is still only a very limited degree of consensus in the academic world on the equity premium and risk-free rate puzzles. This lack of consensus is well illustrated by the views of the two original authors who identified the puzzle in the first place. While Mehra has concluded that the puzzle requires some modification of the underlying theoretical model (as for example, in Constantinides, Donaldson, and Mehra (2002), *op cit*), on the assumption that there is a significant equity premium requiring explanation, Prescott has recently concluded that it is reasonable to treat the equity premium as if it were so close to zero as to be empirically negligible (McGrattan and Prescott (2001)—discussed further in Section 2.6 below).

2.4. Estimating the Common Components from Historic Averages

2.4.1. Background and Caveats

When Mehra and Prescott (1985) first formulated the equity premium puzzle, they used an estimate of the arithmetic risk premium[10] of equities over cash of just over 6%, based on realised returns over the period 1889-1978. This figure was subsequently widely quoted in both popular and academic discussions, and even larger figures have also been quoted, by extending the sample into the bull market of the 1980s and 1990s, during which equities outperformed cash by a wide margin. It is probably fair to say that there is now a reasonable degree of consensus that these initial estimates were almost certainly overstated; however there is less consensus on the extent of this overstatement.

Before proceeding to examine this more recent evidence, some caveats are in order.

First, although it is common to estimate risk premia by using historic average returns, it is worth considering for a moment the problems involved in so doing. A true measure should capture the risk premium that investors are expecting to receive from equities, compared to safer assets, but this must obviously be unmeasurable. The only thing that can be measured is the returns that they have actually received in the past.

It is evident that even over quite long periods, realised returns need not provide any relation to the expected premium. If they did, the experience of the bull market of the 1990s would have implied a risk premium of equities over cash of around 15%, switching to a large *negative* risk premium in the subsequent bear market of the early years of the new millennium. This would be manifestly absurd. There is no evidence that rational investors were expecting to receive such returns in advance. A significant element in the returns they actually received was therefore almost certainly due to expectational errors.[11]

This problem can only be overcome, if at all, by assuming that, if a long enough period

[10] See Section 2.4.2 for a discussion of alternative averaging procedures.

[11] It is important to stress that the theory of rational expectations is perfectly consistent with investors making errors, as long as their previous expectations were formed rationaly - ie, using all available information, as efficiently as possible For arguments that expectational errors have had a significant impact, see Dimson, Marsh, and Staunton (2001a) and Fama and French (2001)

is chosen, pleasant mistakes in predicting returns, such as those of the 1990s, will be offset by unpleasant ones, as more recently. Unfortunately, it is quite possible that historic errors do not always so conveniently average out at zero. And the problem is compounded if, as is often assumed, the underlying thing being measured is itself not constant.

Second, in assessing historic averages of the equity premium and the "safe" return, it is important to treat them consistently. Thus, for example, some arguments, examined below, that the historic risk premium overstates the true risk premium due to one-sided inflation surprises that have depressed the historic safe rate, imply precisely offsetting errors in the two elements, and thus should not affect the estimate of the overall cost of equity capital.[12] For this reason, in what follows we examine evidence on both simultaneously (or, equivalently, on the premium and the return on stocks).

Third, given the highly integrated nature of modern capital markets, there is no reason to restrict attention to data solely from the UK market, and on occasion to do so can actually produce distortions. There is a strong *a priori* case for treating the common components of the cost of equity as being determined in world markets; and, as we shall show, there is also a reasonable amount of evidence in favour of this.

Fourth, different studies report different measures of historic average returns. Before proceeding to examine the historic evidence, we thus need to digress briefly in dealing with the (often quite important) quantitative differences between these alternative measures.

2.4.2. Geometric vs Arithmetic Averaging, and the Role of Return Predictability

2.4.2.1. Definitions Let R_{jt} be the return on some financial asset, defined by

$$1 + R_{jt} = \frac{P_{jt} + D_{jt}}{P_{jt-1}} \tag{2.5}$$

[12] As a counter-example, Giles and Butterworth (2002) in their submission on behalf of T-Mobile, attempt to have their cake and eat it, by basing equity premia estimates on long historic averages in which the safe rate may have been underestimated, but basing safe rate assumptions on recent data, thus generating a high implied cost of equity capital.

where P_{jt} is the price of the asset, and D_{jt} is any dividend or other income generated by the asset.

Standard theory requires that the appropriate measure of any given return used in deriving the cost of capital should be $E(R_{jt})$, i.e. the true arithmetic mean.[13] This requirement holds whatever the nature of the process that generates R_{jt}.

In contrast, historical studies frequently quote two alternative, but closely related, measures. One is what is often rather loosely described as the "geometric mean"; the other is the arithmetic mean of the logarithmic return.

It is very commonly assumed that returns are lognormal (inter alia, this deals with the skewness of returns, and rules out returns of less than –100 %), i.e., letting lower case letters define log returns,

$$r_{jt} \equiv \log(1 + R_{jt}) \sim N(E(r_{jt}), \sigma(r_{jt})) \tag{2.6}$$

but standard properties of the lognormal distribution imply that

$$1 + E(R_{jt}) = exp\left(E(r_{jt}) + \frac{\sigma^2(r_{jt})}{2}\right) \tag{2.7}$$

implying, to a linear approximation the following relationship between the arithmetic mean return, and the arithmetic mean log return

$$E(R_{jt}) \approx \log\left(1 + E(R_{jt})\right) = E(r_{jt}) + \frac{\sigma^2(r_{jt})}{2} \tag{2.8}$$

The latter is in turn closely related to the "geometric mean" return,[14] $G(R_{jt})$ defined by

$$1 + G(R_{jt}) = exp(E(r_{jt})) \tag{2.9}$$

But, again, to a linear approximation,

$$G(R_{jt}) \approx \log(1 + G(R_{jt})) = E(r_{jt}) \tag{2.10}$$

[13] See, for example, the treatment in Copeland and Weston (1992) Chapters 7 and 13.

[14] More strictly defined as the compound average return, constructed in the data as the geometric average of $1 + R_{jt}$, minus one. But, for brevity, we follow standard practice in referring to this magnitude as the geometric average.

with the omitted terms biasing the approximation downwards by a fairly trivial amount.[15] The geometric mean return is a natural metric of returns viewed from the perspective of an investor: an investment with a positive geometric mean return will grow over time.

Thus, as an illustration, suppose that the volatility of log returns is 0.2,[16] a rough ballpark figure for a range of equity markets, according to figures from Dimson, Marsh, and Staunton (2001a) cited below in Section 2.4. The implied difference between the arithmetic and log (hence geometric) means will approximately equal $0.2^2/2 = 0.02$, or two percentage points. The difference between the two measures of mean returns is therefore non-trivial. For higher estimates of volatility, the gap rises sharply: *eg*, for volatility as high as 0.3, the gap increases to approximately $0.3^2/2 = 0.045$, or 4 1/2 percentage points.

In contrast the difference between mean log returns and the geometric mean return is very small: if, for example, $E(r)$ =0.06, the compound average return will equal 6.18%.

Note that the relationship between geometric and arithmetic average returns implies the somewhat counter-intuitive result that, in principle, an asset may have a negative geometric mean return (ie, over long periods, an investor in the asset will lose money), but at the same time a positive arithmetic mean return.

2.4.2.2. The Impact of the Choice of Time Period One issue that is frequently raised is whether the choice of time period can affect the estimate of $E(R_{jt})$, since arguably regulation should be framed in terms of fairly long periods. It turns out that the crucial issue is, as in many contexts, whether returns are predictable or not. Thus, consider the expectation of the five year return, under the same assumptions of log-normality

$$1 + E(R_{jt}(5)) = exp\left(5E(r) + \frac{\sigma^2(R_{jt}(5))}{2}\right) \tag{2.11}$$

where $\sigma^2(R_{jt}(5))$ is the variance of (non-annualised) log returns over five years. Annualising

[15]The mean log return is an unbiased measure of the continously compounded average return.

[16]Note, in passing, that the appropriate measure of volatility, $\sigma(r_{jt})$ is the standard deviation of log returns; in contrast $\sigma(R_{jt})$, the standard deviation of absolute returns will be larger.

the five year return,

$$[1 + E(R_{jt}(5))]^{1/5} = exp\left(E(r) + \frac{\sigma^2(R_{jt}(5))}{10}\right) \tag{2.12}$$

If log returns are unpredictable, then the variance of the five year log return will simply be five times the variance of the one year return, i.e.$\sigma^2(R_{jt}(5)) = 5\sigma^2(R_{jt})$, implying that the change of time period has no impact: the annualised expected five year return is simply the expected one year return.

However, as Campbell (2001b) and Robertson and Wright (2002) have pointed out, if there is predictability of returns, this can significantly lower long-horizon return variances, compared to the random returns benchmark. As an illustration based on annual US data from 1900-2000, Table 2.1 shows estimates derived from Robertson & Wright (*op cit).* The estimated variance of one-period returns is very similar to the illustrative figure above. If instead returns are predicted from a cointegrating vector autorogressive (CVAR) model in which both the dividend yield and Tobin's q have predictive power, the one year ahead (conditional) variance is slightly, but only slightly reduced. However, consistent with much evidence that there is greater predictability of returns at longer horizons, five and ten year return variances are significantly lower than they would be if returns were random.[17]

The implication of these figures is that if they truly capture return predictability, the gap between the arithmetic mean return and the geometric return would fall to only around one percentage point over a five year horizon, and even less over a ten year horizon.

2.4.2.3. So Which To Use? The discussion above shows that the relationship between geometric and arithmetic average returns:

- will only be constant over time if volatility of returns is constant;
- will only be constant across different return horizons if returns are unpredictable.

[17]Note that both sets of estimates also include an allowance for uncertainty with respect to the true mean return (for simplicity the numerical example in the text ignores this issue) this means that in the random returns case variance rises slightly faster than the forecast horizon. But the impact of the adjustment over these horizons is quite small.

	Random Returns	CVAR	Ratio
1 Year	0.044	0.036	0.82
5 Years	0.222	0.118	0.53
10 Years	0.455	0.180	0.39
Source: Calculations based on estimated systems in Robertson and Wright (2002).			

Table 2.1: Return Variances at Different Horizons (Non-Annualised)

Unfortunately, arguments have frequently been presented in the literature that neither of these conditions will hold. There is no doubt that the ultimate aim must be to derive an estimate of the arithmetic mean return, since, as noted above, this corresponds to the theoretically desirable "true" expectation. But if the above conditions do *not* hold, any presumption that, e.g., the arithmetic mean return has been stable over time must, logically imply that the geometric mean return has *not* been stable over time; and vice versa. There is no clear-cut empirical evidence, that we are aware of, that distinguishes between these two characterisations of the data; indeed, given the degree of uncertainty in historical averages, it would be surprising if there were. Eminent academic economists have come down on both side of the fence. Thus e.g., Campbell and his various co-authors typically assume lognormality, as in (2.6), and hence stability of the mean log return and the geometric average, as implicitly, do Dimson *et al.* In contrast, eg, Fama and French have, in various papers, worked on the assumption that the arithmetic mean return is stable.

Our (not very strong) preference would be to side with Campbell, since the assumption of lognormality of returns is consistent with the feature of financial returns that they cannot fall below -100%, but are unbounded in the opposite direction. But given the absence of a clear consensus on the best way to model the underlying properties of returns, the only clear-cut recommendation must be to deal consistently with the difference between the two averaging methods, to be precise in noting which estimate is being used in any context, and to be aware of the potentially significant differences between the two.

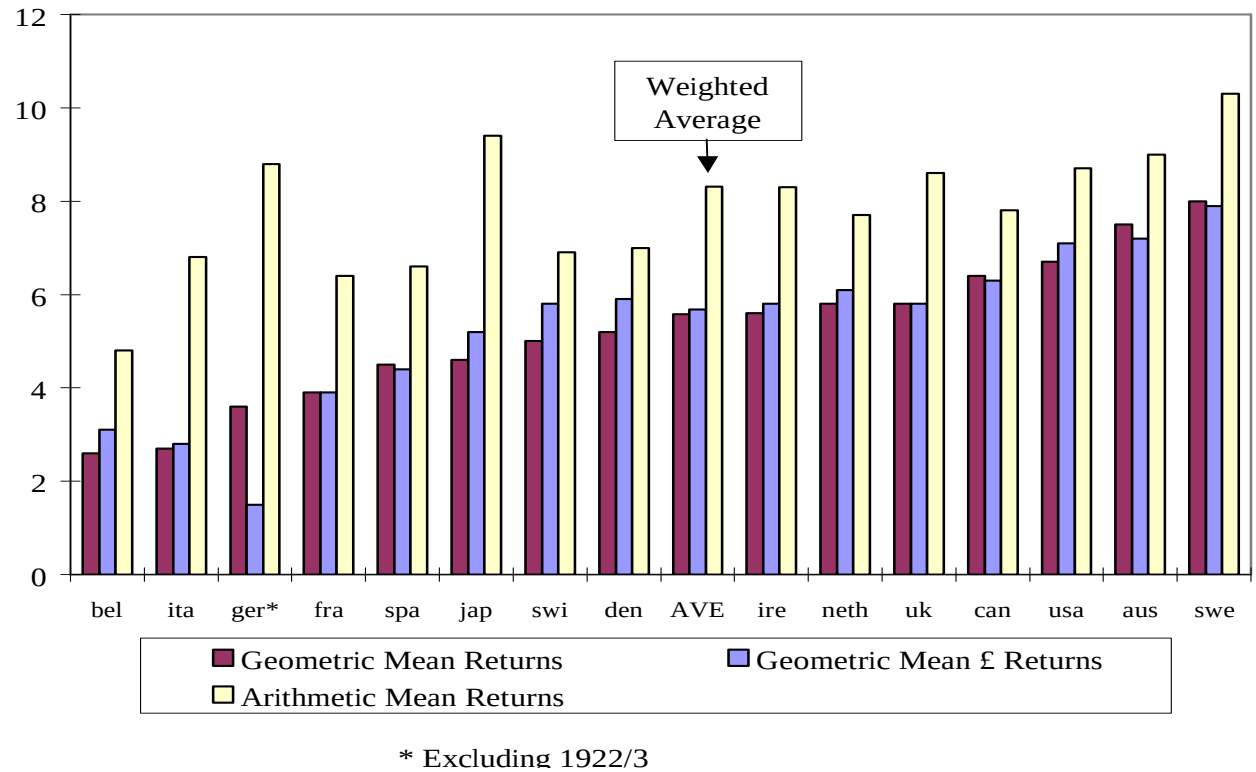


Figure 2.2: Equity Returns, 1900–2000

2.4.3. International Evidence on Historic Returns and Premia

The advent of the LBS/ABN/AMRO database (Dimson, Marsh, and Staunton (2001a), Dimson, Marsh, and Staunton (2001c)) has generated an abundance of new evidence on the common components. Previous estimates had been very dependent on the US, and to a lesser extent the UK markets. While the US market still provides the only consistent source of data for very long runs (i.e., more than a century's worth) of data (discussed below, in Section 2.4.4) a number of authors had cast doubt on whether its experience was truly representative. However, until recently, mult-country evidence such as that of Goetzmann and Jorion (1999) was equally plagued by problems of sample inconsistency, and, crucially, absence of data on total returns (as opposed to just capital appreciation).

Figures 2.2 and 2.3 summarise the key features of the international evidence, based on returns from 1900–2000, from Dimson, Marsh, and Staunton (2001a). The key features to

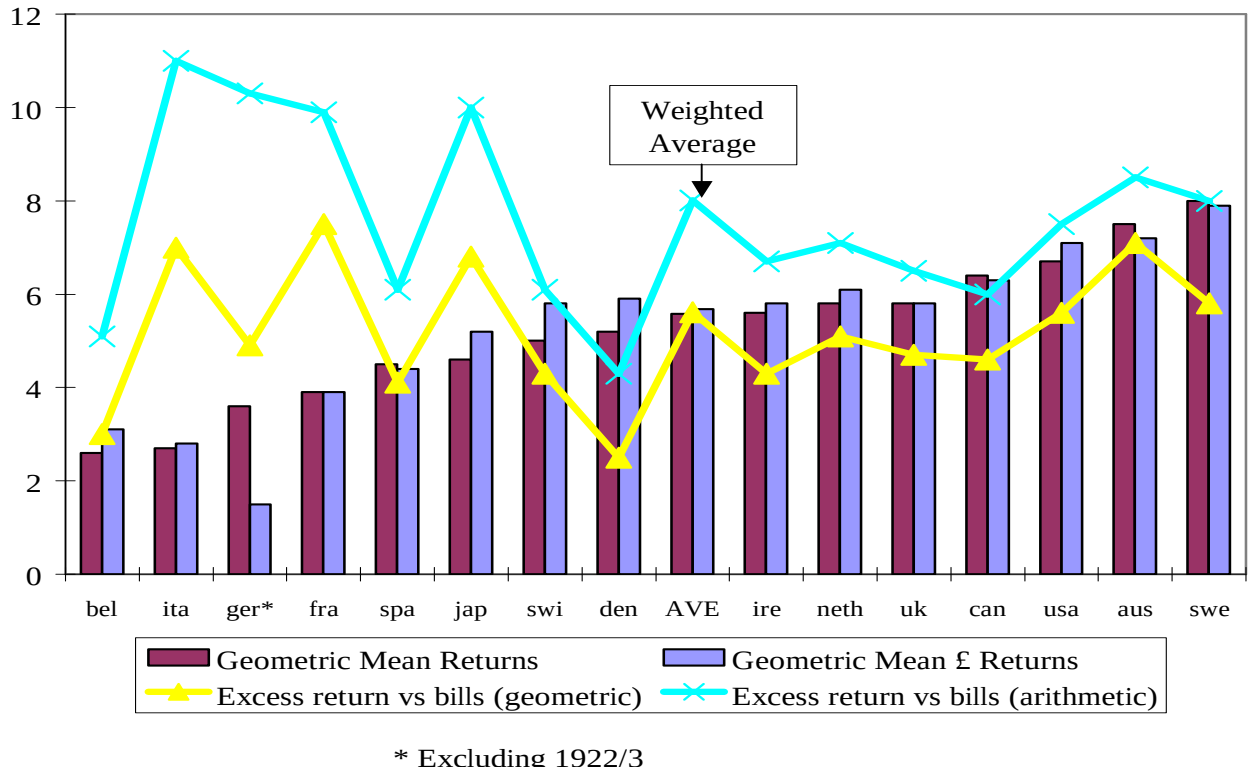


Figure 2.3: Equity Returns and Premia, 1900–2000

note are:

- While there has been a reasonably wide range of experience in the countries covered, the range of historic mean equity returns is not actually all that wide: all but three countries had geometric (ie, compound) average stock returns in the range of 4% to 8%, and arithmetic average returns in the range 6% to 10%.

- In common currency terms, average returns follow a very similar pattern (implying indirectly that purchasing power parity was reasonably close to holding).

- The experience of the UK, in terms of mean returns, has been close to average;[18] that of the US somewhat better than average (though not as markedly so as had been suggested by the earlier work of Goetzmann and Jorion (1999)).

- The difference between geometric and arithmetic average returns is always significant, and is generally distinctly larger for the poorer performing countries. The difference reflects the fact that the standard deviation of returns for the poorer-performing countries tended to be higher than for the better-performing countries. This illustrates the important role geometric vs arithmetic averaging can play, as discussed in Section 2.4.2. In the case of Germany, in particular, equity returns were so volatile that, while it displayed a relatively poor performance in terms of geometric average returns, in terms of arithmetic average returns its performance appears relatively good. In this particular case, we would regard the relative ranking of geometric returns as more representative.

- There is a very similar range of values for the observed geometric risk premium,[19] which is in turn almost unrelated, across the cross section, with the return itself. In a number of countries, by implication, there were common shocks to both stock returns and returns on the "safe" asset (which was of course not at all safe at times of inflation or, especially, hyper-inflation - a theme we revert to below, in Section 2.5.2).

[18] The averages shown in the chart are cross-sectional weighted averages, where each countries weight is the average of their share in market value in 2000, and their share in GDP in 1900 (the best proxy Dimson et al have for market weights at the start of their period). Equally weighted average returns are rather lower, due to the impact of a number of relatively small poor-performing countries.

[19] We follow convention in referring to these estimates as observed risk premia, although they are of course simply average excess returns.

- The arithmetic risk premia are slightly *negatively* correlated with average returns over the cross section: the difference reflects, again, the fact that the standard deviation of returns for the poorer-performing countries tended to be higher than for the better-performing countries.

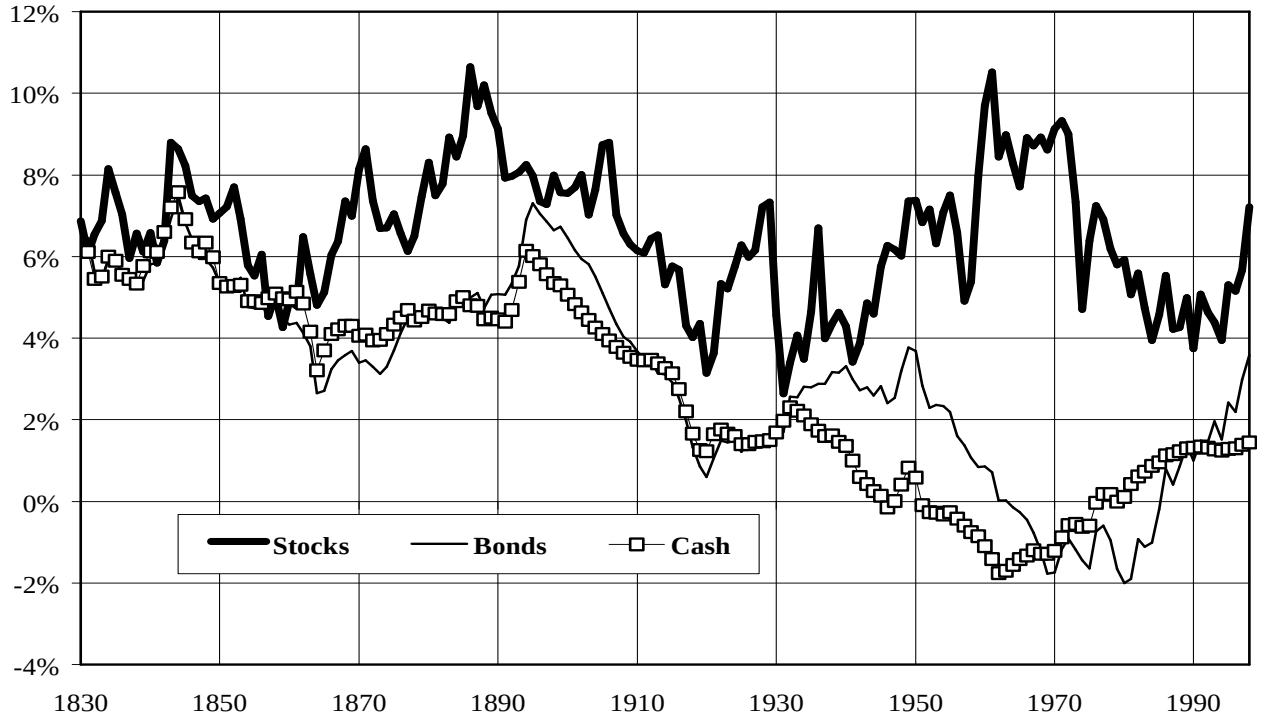
- Cross-sectional differences in mean returns and excess returns are thus not explicable by differences in volatility. Indeed the correlation goes the wrong way, implying an that the range of associated "Sharpe Ratios" (excess returns divided by volatility) is distinctly wider than the range of excess returns. Hence, unless representative investors in different countries had very different degrees of risk aversion, *and* markets were extremely segmented (neither of which is very plausible) this is further indirect evidence of significant expectational errors that did not cancel out.

2.4.4. Evidence from Two Centuries of US Data

Siegel (1998), using two centuries' worth of data for real returns on stocks, bonds and bills in the US, has forcibly argued for the apparent stability of real returns on stocks, both in absolute terms, and relative to competing assets. Smithers and Wright (2002) christen the apparently stable geometric mean stock return "Siegel's Constant".

Figure 2.4 summarises the empirical basis for "Siegel's Constant", by comparing rolling compound average real returns derived from data for the entire two hundred year period (the use of long-period rolling returns is not essential, but helps clarify the graphical presentation). Thus the returns for the dates shown are the compound (or "geometric") average returns on an investment made thirty years earlier.

The chart helps to provide some additional insight into the large estimates of the equity premium for a number of countries based on data for the twentieth century, discussed in Section 2.4.3. While the thirty year stock return has moved within a relatively narrow range, around what does indeed seem to be a stable mean value, real returns on both bonds and bills appear much less stable. The low estimate of the mean real interest rate (i.e., bill return) for the twentieth century, of only around 1%, for example, is seen to result from a pattern of real rates that began and ended the century at significant positive levels, offset by



Source: Calculations using data from Siegel (1998)

Figure 2.4: 30 Year Real Returns on Stocks, Bonds and Cash Since 1830

the period of zero, or negative real rates in the thirty years or so after the second world war. In contrast, returns on both bonds and bills were significantly higher in the 19th century.

Note that if "Siegel's constant" really were constant, or at least close to being so, this would offer some empirical support for the argument, discussed in 2.2, that the equity premium puzzle and risk-free rate puzzles are one and the same, since there is no obvious conflict between observed mean stock returns and the predictions of theory.

Smithers & Wright (*op cit*) calculate confidence intervals for the mean log stock return[20] over the entire two century period, using different non-overlapping horizon returns. Table 2.2 reproduces these results. If (log) returns were strictly random, then the horizon adjustment should make no difference, but predictability implies non-randomness, and hence that the confidence intervals calculated for one-year returns may be distorted. Over longer horizons, this non-randomness is less relevant. Although the number of observations is reduced markedly, the associated reduction in variance more than offsets this, such that the confidence interval for the mean return is narrower if long-horizon returns are used.

	95% Lower Bound	Mid-Point	95% Upper Bound
¿From annual data	0.0426	0.067	0.0924
¿From non-overlapping 10 year returns	0.045	0.066	0.086
¿From non-overlapping 20 year returns	0.038	0.063	0.088
¿From non-overlapping 30 year returns	0.049	.063	0.077

Table 2.2: Confidence Intervals for "Siegel's Constant" (the mean log stock return)

Dimson, Marsh, and Staunton (2001c) dispute the concept of "Siegel's Constant" being a global phenomenon, in the light of the variation across countries in their sample, as shown in Figure 2.2.[21] It is indeed almost certainly the case that, as proposed by Goetzmann and

[20] Assuming lognormality of returns. Given the relationships examined in Section 2.4.2, implied figures for the geometric average would be virtually identical.

[21] Although it should be noted that their arguments are mainly framed in terms of the equity premium, rather than the equity return.

Jorion (1999) and others, the experience of the US market may overstate the true world expected stock return, due to a form of survivorship bias. Even over two centuries there may be an impact of non-offsetting expectational errors biasing up the US mean return, since the extent of the success of the US economy could not have been predicted. Dimson *et al*'s data, summarised in Figure 2.2 showed that the geometric average world return over the twentieth century was indeed around a percentage point lower than that of the US.[22]

Such a difference, it should be noted, lies well within the confidence interval shown in Table 2.2, and is also consistent with other adjustments to the mean return, discussed below (in Section 2.4.5) that take into account the possible distorting effects of the exceptional returns during the 1990s.

With or without the adjustment to the point estimate, the relatively narrow confidence interval for the mean real stock return are worth bearing in mind, given the point noted in Section 2.1.2, that, for firms with beta reasonably close to unity, and with relatively low gearing, cost of capital calculations are dominated by the assumed cost of equity, with only a relatively small role for the safe rate, and hence for the equity premium.

It is not possible to use Siegel's dataset to derive plausible confidence intervals for the real returns on bills and bonds, and hence for the equity premium. The problem, is that, as Figure 2.4 shows, there is far less evidence of stability in these magnitudes, implying that any confidence interval derived on the assumption that their "true" values are constant is almost certainly mis-specified.

One possible explanation for this apparent instability, noted by a number of authors is that there have been major expectational errors in inflation – the key determinant of real returns on these competing assets. This has of course been even more important in a number of other countries in the Dimson *et al* sample, that suffered high, or even hyper-inflation during the twentieth century. It might be that the true underlying premia were much more stable, or even constant, once account is taken of such errors.

Pickford and Wright (2000) attempt to uncover the mean underlying premia from the Siegel dataset, by calculating mean returns over periods in which expectational errors were

[22]The gap in terms of arithmetic averages was narrower, since non-US markets had higher volatility.

probably reasonably close to zero.[23] There is a slight decline in the implied mean stock return between the 19^{th} and 20^{th} century (since data since 1974 are excluded in order to avoid the possible distorting impact of the 1990s boom). In contrast, the implied expected return on cash over the 20^{th} century is raised by around a percentage point from its actual average value, and the estimates of expected bond yields appear remarkably stable across two centuries.[24]

	Cash	**Bonds**	**Stocks**	**Geometric Premium Over Cash**
All relevant data	3.44%	4.16%	6.61%	3.17%
Only 20^{th} Century	1.76%	4.22%	6.01%	4.25%

Table 2.3: "Expectations-Neutral" Compound Average Real Returns, and the Equity Risk Premium.

The combination of a lower estimated stock return and a higher estimated cash return over the twentieth century suggests that, using these estimates, the realised geometric risk premium over the same sample, of just under 6%, is almost certainly overstated, possibly by as much as 2 or 3 percentage points.

2.4.5. Evidence on Mean Stock Returns Allowing for Predictability and/or "Over-Valuation" during the 1990s

Another possible source of upward bias in the mean stock return that has been put forward is that the rise in the stock market during the 1990s was sufficiently strong as to affect even

[23]Inflation expectations are proxied by a univariate time series model, derived from priors consistent with data up to that point. Thus, throughout the 19^{th} century, and well into the twentieth, inflation expectations are assumed to be mean zero. On the basis of annual data, this null could not have been rejected on standard tests until after the second world war.Cash and bond returns are calculated as compound averages over samples in which inflation expectations were realised, weighted by sample size (hence some data are excluded from averages). Stock returns (that are arguably less affected by one-side expectational errors) are calculated as trough-to-trough compound averages (hence data after 1974 are excluded).

[24]There is still a distinct difference in the implied real return on cash, but this may well reflect data problems. Banks were more dangerous places to invest money than government bonds, up until the mid-1930s. There is therefore a problem, during the 19th Century when short-term government paper was seldom available, in estimating the risk free return on cash.

long-period averages. This point is noted in the work of Fama and French (2001) discussed below, although only in the context of the bias over the (relatively short) post-war sample. Giles and Butterworth(2002) illustrate this for the UK by assuming that, in line with earlier analysis of Smithers & Co, the UK market was twice overvalued at the end of the century. Spread over the course of a century (and assuming the market to have been fairly valued at the start of the period), this would have raised the compound average return rate by $2^{1/100}$-1$\approx$0.7% per annum: a non-trivial adjustment.

Robertson and Wright (*op cit*) show that a similar result can be derived for the US return, from econometric estimates that allow for the joint determination of stock prices, dividends and the capital stock. If "Tobin's q" (approximately given by the ratio of the stock price to the capital stock per share) and the ratio of stock prices to dividends per share are both mean-reverting, then over long samples, all three series must grow at the same rate, and data on all three series thus provide, effectively, a pooled estimate of this common growth rate.[25] This in turn can be shown to imply an estimate of the mean stock return (using Campbell & Shiller's (1988) approximation). Table 2.4, reproduced from Robertson & Wright, shows that this estimate is pulled downwards, compared to the mean realised log return, given that prices at the end of the sample rose so far out of line with the other series.

The table also shows that, conditional upon the model being correctly specified (i.e., assuming that both ratios do really mean-revert), the resulting estimate of the mean log stock return is very much better determined than if the mean return is estimated from returns alone (cf. the results in Table 2.2): the 95% confidence interval for the point estimate lies between 0.057 and 0.073 – i.e., only barely includes the realised mean return.

	From returns alone	From system
Point Estimate	0.073	0.065
Standard Error	0.021	0.0042

Table 2.4: Alternative Estimates of the US Mean Log Stock Return, 1900-2000

It should be stressed, however, that the predictability of returns that underpins this ap-

[25]There is a strong parallel with the results of Fama & French (2001), discussed below, except that they do not exploit the pooled estimation procedure to improve the precision of their estimates.

parent degree of parameter precision would certainly not command unanimous agreement amongst financial economists. While there is a very large literature that assumes predictability of returns, in recent years there has also been a steady flow of papers that have cast doubt on this evidence, as due to data mining.[26]

2.4.6. Learning from the UK Experience: "Siegel's Constant" and the Consumption CAPM Revisited

We have seen that, while there is strong evidence, presented in Sections 2.4.4 and 2.4.5, that the mean stock return in the extremely well-documented US market appears quite well-determined, Section 2.4.3 showed that it is almost certainly an overestimate of the mean historic world return. This is unsurprising, given the US's relative degree of success over the historic sample covered by available data.

While we advocate strongly the use of international data in assessing the true cost of equity capital, it is worth noting, that, viewed in that context, the experience of the UK, that appears close to the international average, may possibly also be more representative on *a priori* grounds. Arguably the macroeconomic environment in the UK was least subject to surprises over the 19^{th} and 20^{th} centuries. In particular it has been well-documented that the UK's growth of GDP per capita has been remarkably stable, at around 2%, for at least the past two centuries. In contrast, most other countries have had periods of distinctly more rapid growth.

However, there does seem to be mounting evidence that the rich countries, at least, have, in the postwar era, converged on to a very similar growth path, that turns out to be very much in line with the UK's long-term historic average. While we would make no claim that this reflects any deep significance about the number 2% per annum, it does suggest that an economy that has had this growth rate for a very long time may provide a particularly good estimate of another apparently fairly stable value: "Siegel's Constant", the mean stock return.

[26] See Campbell et al (op cit) and Campbell (2001) for excellent surveys. Goyal and Welch (2002) is a recent example of an attack on the predictability literature.

We noted in Section 2.2 that, while the "Consumption CAPM" has major problems explaining the observed equity premium, it is not obviously inconsistent with the observed mean stock return itself. That model suggests that, apart from "deep" risk aversion and intertemporal preference parameters, and covariances, the key element determining equilibrium returns is the growth rate of consumption. This suggests that, if "deep" parameters are reasonably stable, an economy like the UK, that has had particularly stable consumption growth, in line with what now appears to be the international norm, may be a particularly good place to look for evidence of stable stock returns.

This conclusion is if anything reinforced by the observation (suggested both by the estimated "expectations-neutral" stock returns shown in Table 2.3, and from other evidence, such as that of Fama and French (2001), discussed below) that US returns in the twentieth century may have been somewhat lower on average than in the nineteenth, and may also indeed have fallen through the twentieth century, to levels not dissimilar from the UK geometric average return of around 5 3/4%.

2.5. Forward- Vs Backward-Looking Approaches to the Common Components

2.5.1. Forward-Looking Adjustments to Historic Returns and Premia

We have already noted that to treat historic average returns as necessarily equal to true underlying expected returns is naïve. There is however a need to distinguish between trying to identify what the past was really like (eg, adjusting for possible one-sided inflation errors, as described in Section 2.4.4), and arguing that, on a priori grounds, the future must be different from the past. The latter is a much riskier enterprise, since by definition such claims cannot be based on any data.

Some possible adjustments that have been proposed have been

- Dimson, Marsh, and Staunton (2001a) propose that arithmetic premia should be adjusted downwards to reflect forward-looking assessments of volatility. To the extent that this reflects clear distortions in the historic record (eg, extreme volatility during

hyper-inflations) this is almost certainly valid. But to the extent that it embodies the assumption that the world is a safer place, this approach is on distinctly less firm ground. There is indeed a reasonable amount of evidence that macroeconomic aggregates like GDP became more stable in the second half of the twentieth century. But, at least in mature markets, the evidence that *stock markets*, as opposed to the rest of the economy, have got much safer, is distinctly weaker. In economies that escaped major disruption, such as the UK or the US, there is little or no evidence of a decline in stock return volatility.[27]

- A much more radical approach, also proposed by Dimson *et al*, is to infer that the equity premium *must* have permanently fallen from the observed fall in the dividend yield. The problem with this argument is that it is driven entirely by the rise in the market during the 1990s. It is certainly a logically possible justification for the high market, but the only evidence for it is the level of the market itself (see Sections 2.6 and 2.5.3 below for further discussion of interpretations of recent experience). It is in distinct contrast to the approach of Fama and French (2001) discussed below.

- Another argument used by Dimson *et al* is that trading costs of forming diversified portfolios have fallen. At the same time, however, the proportion of the population investing indirectly in the stock market has risen enormously. The rise of 3^{rd} party investment, via pension funds, etc, may quite possibly have increased principal-agent type costs for the *average* investor. There is certainly evidence that the costs of 3^{rd} party investment are distinctly non-trivial: the table below summarises data on the costs of retail investing from James (2000) study for the FSA. Thus the case for lower trading costs does not appear clear-cut.

- It is frequently claimed that financial liberalisation has eased credit constraints significantly, and that this may result in a lower premium. This argument has been put forward by e.g., Heaton and Lucas (1999). We noted in Section 2.3 that in the model of

[27] Fama and French (2001) note that observed volatility in the US market fell slightly in the postwar period, but the fall was well within the confidence intervals associated with an assumed constant rate of true volatility. Nor is there any obvious downward trend in "implied volatility" estimates derived from options prices (see Smithers & Wright, 2000, Chapter 30).

	UK Actively Managed	US Actively Managed	UK Index	US Index
Explicit Costs	1.4%	1.45%	0.98%	0.45%
Implicit Costs	1.31%	0.92%	0.88%	0.41%
Total Implicit and Explicit Costs	2.71%	2.37%	1.86%	0.86%
Upfront Charge/Bid-Offer Spread	5.34%	1.34%	2.66%	0.27%
Total (10 year average hold)	3.24%	2.50%	2.13%	0.89%
Source: James (2000)				

Table 2.5: Retail Investing Costs

Constantinides *et al* (*op cit*), credit constraints may indeed help to explain the historic equity premium, and a reduction in such constraints would indeed imply a lower equity premium. However, this would result from a marked *rise* in the return on safe assets, not from a fall in the return on equities.

Overall, the impact of such hypothesised shifts is very hard to quantify. And, crucially, it is worth stressing that, as in the last example, if any such effects did imply a fall in the equity premium, this need not necessarily imply a fall in the equity *return*: i.e., it could just as easily imply a rise in the safe rate. Indeed, in the next section we discuss arguments that on *a priori* grounds this might appear more likely.

2.5.2. Forward- Vs Backward-Looking Measures of the Risk-Free Rate

We noted in Section 2.1.2 that, for most firms with CAPM beta not too far from unity, the risk-free rate (and hence the equity premium) plays a more minor role, compared to the estimate of the market return itself: or equivalently the return on an equity with a beta of precisely unity. This is fortunate, since not only are historic average values of the risk-free rate hard to rationalise with theory (as we saw in Section 2.2); but they are also much harder to rationalise with that average value being stable over time (as we saw in the context of the US experience in Section 2.4.4).

Figure 2.5 shows that appealing to international evidence (taken, again, from Dimson,

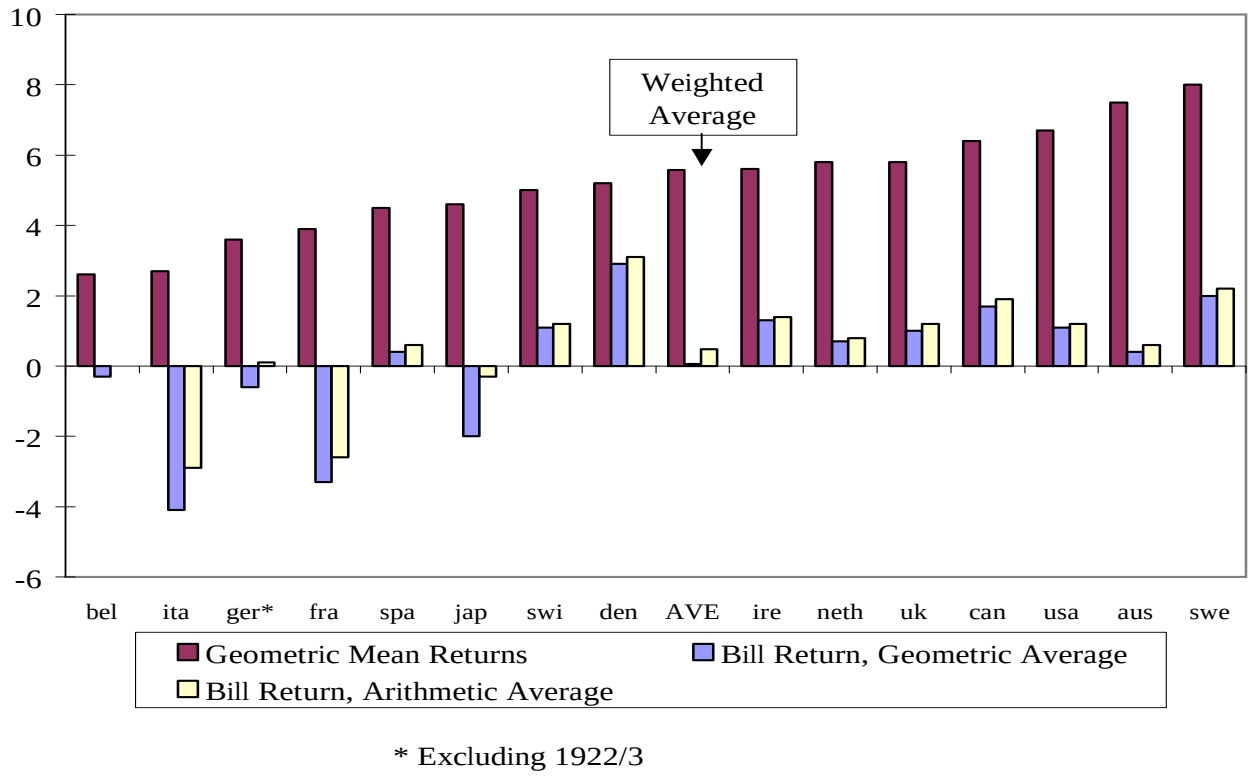


Figure 2.5: Returns on Equities and Bills, 1900-2000

Marsh, and Staunton (2001a)) does not provide much assistance. The range of mean real returns on short-term bills over the twentieth century is actually wider than the range of equity returns over the same set of countries, with a number of countries having experienced negative real interest rates. The weighted average real interest rate for the Dimson *et al* set of countries over the course of the twentieth century was essentially zero. As noted above, however, arguably all of the countries in the sample suffered one-sided inflation surprises over this period; and the particularly poor performers were typically countries that at some point in the sample suffered particularly badly in this respect.[28]

In the absence of clear evidence of a stable mean over long samples, there may be better arguments for a forward-looking approach in setting the risk-free rate. This approach is aided, as in the case of the cost of debt (discussed in Section 2.1), by the fact that at least

[28]Note that data for Germany exclude the hyperinflationary years 1923/4; if these were included even the average real interest rate for Germany would go right off the scale.

current values of the risk-free rate can be observed directly from the data. While short-term interest rates are still subject to inflation risk, the advent of independent central banks has done a great deal to reduce inflation uncertainty, so at least over short horizons the forward-looking real safe rate can be estimated with considerable reliability.

Of course, if assumptions are made about the safe rate over reasonably long horizons, it is not sufficient simply to take a snapshot of whatever the current short rate happens to be. Various forward-looking alternatives are available:

- Market expectations of future short rates can in principle be inferred from futures prices. These are however known to be biased predictors of actual future spot rates, since they arise by arbitrage across different maturities in the yield curve - they thus also include an implicit impact of term premia. Adjustments allowing for this bias are however relatively straightforward to calculate.
- As an alternative, or, more likely, as a complement to this approach, non-market forecasts of future short rates are readily available, both from private sector consensus forecasts, and from independent forecasting bodies such as the National Institute for Economic Research.
- The availability of data on indexed bond yields also implies that, at any given horizon, there does now exist a perfectly safe asset, in the form of an indexed bond with a maturity equal to the desired horizon.[29] However, information from indexed bonds should be treated with some caution. First, there are known distortions due to tax treatments. Second, the relatively short sample over which they have been available implies that there is relatively little evidence on their true expected return (assuming this to be stable). The era since the advent of indexed bonds has largely been one of falling inflation: as a result realised returns have been poor, while yields have been fairly high. Third, and crucially, estimates derived in this way are inconsistent with estimates derived over longer samples in which such assets were not available.

While the above methods, if carefully applied, are likely to give a reasonable estimate

[29]Note however that indexed bonds of an given maturity are not safe at intermediate horizons, since they are subject to market real interest rate shocks.

of current consensus views on the real safe rate, it must be acknowledged that, given the difficulties in interpreting the historical data, discussed above, such consensus views may well exhibit some degree of instability over time. Realistically, most available long-term forecasts of real short rates are likely to be driven by assumptions about equilibrium real interest rates drawn from relatively short samples. Thus, for example, the common assumption in discussions of monetary policy along "Taylor Rule" lines,[30] that the mean real interest rate should be of the order of 2 1/2%, is largely driven by experience since the 1980s.[31]

Finally, we discussed in Section 2.2 the argument that the "equity premium puzzle" and the "risk-free rate puzzle" are essentially the same puzzle. The empirical evidence of the preceding sections provides some support for this argument. That being the case, if the equity premium *were* to fall in the future, it would seem more likely that the safe rate (both hard to explain and apparently unstable in the data) should rise towards the more stable, and more easily explicable stock return, than that the latter should fall. There has indeed been an observable tendency for real safe rates to drift upwards from the second half of the twentieth century onwards - visible both in the US data (see Figure 2.4) and, as Dimson, Marsh, and Staunton (2001a) show, in a wide range of other countries.

2.5.3. Inferring Equity Premia and Expected Returns from the Dividend Discount Model

A number of authors have inferred the desired equity return and/or the equity premium from variants of the dividend discount model (e.g., Blanchard (1993); Wadhwani (1999); Heaton and Lucas (1999); Fama and French (2001)).

Thus, Fama & French use the simplified (i.e., constant growth rate) Gordon (1962) model to derive the relationships:

$$P_t = \frac{D_{t+1}}{R - G} \tag{2.13}$$

implying

$$R = \frac{D_{t+1}}{P_t} + G \tag{2.14}$$

[30] Taylor (1993). See Clarida, Gali, and Gertler (1999) for a helpful survey.

[31] An estimate of around this size is also consistent with estimates from Pickford et al (*op cit*) that the "expectations-neutral" return on the safe asset in the USA (as in Table 2.3) over the sample 1976-1998 (a period in which estimated expectational errors were close to zero) was just over 2%.

where (following Fama & French's terminology) R is the discount rate (required return on stocks), and G is the growth rate of dividends.

A disadvantage of the specification in (2.13), noted by Heaton & Lucas (*op cit*), Dimson, Marsh, and Staunton (2001c) and many others, is that relatively small changes in the discount rate or growth rate can imply massive shifts in the implied stock price. But as a corollary, if the objective is to estimate R, even large shifts in the stock price should not greatly change an estimate derived by using some version of (2.14). Fama & French show that estimates of R derived in this way from US data are much more stable, and more consistent with other data (for example, underlying rates of profitability) than those derived by using realised returns.

It should be stressed that Fama & French primarily promote this approach as a method of estimating the stable long-run mean return; indeed, their statistically based arguments for the superiority of this approach are predicated on the assumption that both returns and dividend yields are mean-reverting. As such, their approach is unobjectionable in principle, and in practice does not produce very different answers for underlying US stock returns or premia from those derived from long historical averages, especially if the latter are adjusted for one-sided errors, as discussed in Section 2.4.4 above; or for mis-valuation and/or predictability, as discussed in Section 2.4.5. Table 2.6 summarises their long-run estimates:

	Real Stock Returns		Excess Returns	
	"Gordon Estimate"	Realised	"Gordon Estimate"	Realised
1872-1999	6.88	8.97	3.64	5.73
1872-1949	7.79	8.10	3.79	4.10
1950-1999	5.45	10.33	3.40	8.28
Source: Fama and French (2001)				

Table 2.6: Estimated Arithmetic Mean Returns

Fama & French note that, for comparability with realised returns, the Gordon estimate should be raised by roughly 1.2% a year, due to the lower variance of dividend growth, compared to stock price changes.[32]

[32]For reasons outlined in Section 2.4.2.

Fama & French also seek to explain the large gap between the two competing estimates of returns and excess returns in the post-war period. In accounting terms, the explanation is straightforward: there was a very large fall in the dividend yield (and hence much more rapid growth of prices than of dividends) over this period, thus the contribution of capital appreciation to realised returns was much greater than that of dividend growth to the Gordon estimate.

Equation (2.13) suggests two possible explanations for this shift: either R must have fallen, or G must have risen.

At the height of the boom this latter explanation was put forward very frequently by defenders of the then level of the market. Fama and French dismiss this argument out of hand on empirical grounds, however, given the very weak degree of predictability of dividend growth.[33]

Fama and French conclude that the only possible explanation is that, at the end of their sample, the expected stock return and the equity premium that a rational investor could expect must have been exceptionally low in historical terms. They calculate that, assuming the dividend yield to remain constant from 1999, the expected real return must have been only just under 3%, and the equity premium (compared to short-term bills) less than a percentage point. They note, however, that even this figure may be overstated, since, if dividend yields are expected to mean-revert (which they assume) expected growth of dividends must be greater than expected stock price capital appreciation: thus the Gordon estimate will overstate the true return at times of low dividend yields.

As firm believers in efficient markets, Fama & French do not take this argument further, - perhaps unsurprisingly, since, for reasonable history-based estimates of the rate of adjustment of dividend yields, this could very easily lead them to an implied equity premium at the end of their sample that was significantly negative, and thus in line with the views expressed

[33]They might also have added a considerable weight of theory to this dismissal, since, as noted by a number of authors (e.g., Kiley (2000); McGrattan and Prescott (2001); Campbell (2001a); Smithers and Wright (2000)) general equilibrium considerations would suggest that any such hypothetical rise should if anything generate a *fall* rather than a rise in the market. The essence of this point is that, in line with the analysis of Section 2.2, rational forward-looking consumers, aware of higher growth rates in the future, would attempt to bring forward their consumption to today. In general equilibrium this is impossible; hence capital markets would only be equilibrated by a *rise*, not a fall in expected returns.

by, *inter alia*, Campbell and Shiller (1988); Wadhwani (1999); Shiller (2000) ; Smithers & Wright (2000); Robertson & Wright (2002) that the only data-based prediction at the end of the 1990s was that real stock returns, and *a fortiori*, excess returns, had a very high probability of being negative.[34]

While this in no way invalidates the methodology Fama & French apply to derive backward-looking estimates of the long-run return and equity premium, it does illustrate the difficulty of using the dividend discount model to derive expected returns and the equity premium in a forward-looking way. Asking what equity premium a rational forward-looking investor would have demanded at the end of the 1990s *may* uncover the true equity premium—but there is no data to substantiate this calculation. It may however simply lead one to the conclusion (so far substantiated by events!) that, at end-1990s prices, a rational forward-looking investor would rationally have got out of the stock market altogether.

2.6. Interpreting the 1990s Boom, and Implications for Cost of Capital Assumptions

Academics and financial analysts alike have been divided on how to interpret the sharp rise in market valuation ratios during the 1990s. As we write, these valuations appear to be unwinding fairly rapidly, with both US and UK markets having fallen to levels only just above half their peak values. But, on most valuation measures, these falls have not yet fully unwound the impact of the earlier exceptional returns. Nor have the issues raised during the boom yet disappeared.

One view of the market highs of the 1990s was as arising from efficient market responses to a permanent fall in the cost of capital (eg, Heaton & Lucas (1999); Glassman and Hassett (1999); and at least implicitly McGrattan and Prescott (2001)[35]). This argument appears at least in part to have been accepted by Dimson, Marsh, and Staunton (2001c). The alternative, more pesimistic view was that such rises (whether or not consistent with market

[34]They allude only very indirectly to this possibility: "the unusually high 1950–1999 returns seem to be the result of low expected future returns (at the end of the sample). *Of course, some might argue that causation goes the other way*" (Italics added).

[35]McGratan & Prescott simply took a snapshot of the market at its peak in 1999, and concluded that it was fairly valued on the assumption that the equity premium was zero.

efficiency) were ultimately likely to prove temporary, with a (possibly prolonged) period of very weak returns, before markets returned to offering returns close to, or equal to historic averages (eg, Cochrane, 1997; Campbell & Shiller, 1998; Shiller, 2000; Smithers & Wright, 2000; – also, implicitly, Fama and French (2001))[36]

The controversy can be encapsulated in two competing hypotheses. The first is that the cost of equity capital has permanently fallen; the second that any such fall is purely temporary.

Kiley (2000) has shown convincingly that, even if the first hypothesis is correct, the impact would be unlikely to be in line with that implied by a naïve interpretation of the dividend discount model. Any permanent fall in the cost of capital would imply a corresponding fall in the return on capital (ie in profitability) in equilibrium (but not necessarily in transition to equilibrium), with the equilibrating factor being the level of capital itself. An important issue for regulation in this scenario would be how best to ensure consistency in the response of regulated industries in the face of any such fall, such that, in any final equilibrium, the relative responses of the capital stocks of regulated and unregulated industries would be optimal.

Any such response would however need to allow for the risk that in fact the second hypothesis may be correct. A lowering of the assumed cost of capital, and hence the target return, for regulated industries, when in fact the equilibrium cost of capital had not fallen would bring significant associated risks that regulated industries may under-invest; may attempt to become lower its beta by being, in effect, unduly risk-averse; or may even ultimately cease operations altogether

A consistent and cautious approach to this dilemma requires an assessment of the relative risk of failing to respond (or responding only with a lag) to a fall in the cost of capital, compared to the risk of responding to a fall that has not actually occurred. We address certain aspects of this issue further in Chapter 5.

[36] A third rationalisation of high valuations based on the assumption of efficient markets was put forward by Hall (2000); but Hall's arguments were not based on an assumed fall in returns, but on the assumption of massive data measurement error: indeed, if anything, he assumed that underlying returns had risen.

2.7. The Common Components: Key Conclusions

- Assumptions about the safe rate and the cost of equity capital must be made consistently.
- Both on *a priori* grounds, and on the basis of evidence, estimates should be formed on the basis of international data, not just from UK experience.
- There is considerably more uncertainty about the true historic equity premium and (hence the risk-free rate) than there is about the true cost of equity capital. From the perspective of the regulators, however, this ranking of uncertainty is fortunate, since the latter is far more important, for firms with risk characteristics not too far from those of the average firm.
- For this reason we regard the standard approach to building up the cost of equity, from estimates of the safe rate and the equity premium, as problematic. We would recommend, instead, that estimates should be derived from estimates of the aggregate equity return (the cost of equity for the average firm), and the safe rate.
- While arithmetic mean returns should be used to proxy for expected returns, these are best built up from a more data-consistent framework in which returns are log-normally distributed, so means should be estimated with reference to mean log returns, or (virtually identically, geometric (compound) averages).
- The longest available dataset, for the US, points to an apparently stable geometric average stock return, over two centuries of data, of around 6–6.5%. This estimate is fairly well-determined (a 95% confidence interval of less than a percentage point on either side of the mean) if returns are predictable; less so (an interval up to 2 percentage points either side of the mean) if they are not.
- International evidence suggests that the US experience was somewhat better than the world geometric average of around 5.5%, which was also close to the UK experience. Although we believe all such estimates should be derived in a world context, there may nonetheless be both empirical and theoretical grounds for regarding the UK's historic cost of equity capital as more typical of the prospective world return.

- The arithmetic mean return may exceed the geometric mean return by as much as 2 percentage points in annual terms (given historical estimates of stock return volatility and an assumption of unpredictable returns). However, if cost of capital assumptions are being made over longer horizons, this may be an over-estimate (possibly by as much as a full percentage point), if either a) returns are predictable; or b) (more dubiously) stock returns in future are likely to be less volatile.

- We are very sceptical of "forward-looking" (and low) estimates of the equity premium and stock returns derived, usually at the height of the boom, from the dividend discount model.

- Our central estimate of the cost of equity capital is around 5.5% (geometric average), and thus 6.5% to 7.5% (arithmetic average). 95% confidence intervals are, at a conservative estimate, of up to two percentage points either side of the point estimates.

- Problems in assessing historic mean values of the safe rate imply that estimates of the future safe short-term rate (that are fortunately of distinctly lower importance for regulators) should probably be derived in a forward-looking way from current rates. However, in so doing, account should be taken of forecast future movements of short-term rates, derived both from market data and published forecasts.

- A commonly used estimate of the equilibrium short-term rate (based on a sample of data from around 1980) is of the order of 2 1/2%. Using this figure, the implied equity risk premium is of the order of 3 percentage points (geometric) and 4-5 percentage points (arithmetic). Given our preferred strategy of fixing on an estimate of the equity return, any higher (or lower) desired figure for the safe rate would be precisely offset by a lower (or higher) equity premium, thus leaving the central estimate of the cost of equity capital unaffected.

- We do not entirely rule out the possibility that the equity premium may fall significantly at some point in the future; but the continuing uncertainty about the premium, both in theory and data, would suggest that, if this does occur, it is more likely to do so through a (further) rise in the safe rate, than through a fall in the equity return.

3. A Comparison of Asset Pricing Models for Regulation

3.1. Introduction

Most assets have some exposure to risk. Common sense dictates that investments that are riskier need to make higher returns to compensate for risk. Various models of risk and return in finance highlight two central points. First, they all define risk in terms of variance in actual returns around an expected return; thus, an investment is riskless when actual returns are always equal to the expected return. Secondly, they all argue that risk has to be measured from the perspective of the marginal investor in an asset, and that this marginal investor is well diversified. It is only the risk that an investment adds to a diversified portfolio that should be measured and compensated.

Despite these common views, differences exist between various asset pricing models as to how to measure market risk. The Capital Asset Pricing Model (CAPM) measures the market risk with a beta measured relative to a market portfolio. Multifactor models measure market risk using multiple betas estimated relative to different factors. In this section, we review four classes of asset pricing model:

1. The Capital Asset Pricing Model.
2. Nonlinear models.
3. Conditional models.
4. Multifactor models.

The objective of the review is to identify the key theoretical and empirical differences between the models, with a view to assessing which approach is most appropriate for the estimation of cost of capital for regulated utilities in the U.K..

It is now well-known that the CAPM has failed to account for several observations about average stock returns (for example, that they are related to firm size, book-to-market equity and a number of other factors). Partly because of the empirical shortcomings of the CAPM, multifactor models (based on the arbitrage pricing theory, APT, of Ross (1976)) have gained in popularity amongst academics and practitioners. More recently still, nonlinear multi-factor models have been developed. The multifactor models have, however, several shortcomings (for example, no adequate test to guard against overfitting of the data). Recent work has shown that conditional CAPMs, which allow the variables in the asset pricing equation to vary over time, may perform better than the standard (i.e., unconditional) CAPM, and as well as multifactor models. These models retain the linear simplicity of the CAPM. They are susceptible, however, to overfitting of the data, and the methodology is some way from being fully-developed and implementable.

This chapter is structured as follows. The Capital Asset Pricing Model is discussed in section 3.2; both the theoretical and empirical foundations for the model. Section 3.3 deals (briefly) with nonlinear asset pricing models; section 3.4 examines models with time-varying parameters. A more extensive analysis of multifactor models is given in section 3.5, dealing particularly with the influential models developed by Fama and French. Section 3.6 concludes, emphasizing the lessons to be drawn for practitioners.

3.2. The Capital Asset Pricing Model

The Capital Asset Pricing Model (the CAPM) was developed 30 years ago by Sharpe (1964) and Lintner (1965). The CAPM was the first apparently successful attempt to show how to assess the risk of the cash flow from a potential investment project and to estimate the project's 'cost of capital'—the expected rate of return that investors will demand if they are to invest in the project.

The algebraic expression of the single-period[1] CAPM is particularly simple:

$$\mathbb{E}[R_e] = R_f + \beta(R_M - R_f)$$

where $\mathbb{E}[R_e]$ is the expected rate of return on equity; R_f is the safe rate of return i.e., the return available on an asset with no risk; beta is the firm-specific beta; and R_M is the expected return on the 'market portfolio' i.e., the expected return from investing in risky assets.

An investor always has a choice to invest in a 'risk-free' investment such as a Treasury bill in which there is no significant risk of losing the investment and where the investment return is fixed. This is called a safe asset. Alternatively, an investor can enter into more risky investments. In order to be persuaded to do so, (s)he will require an additional return. The equity risk premium can be thought of as the required excess return on a portfolio made up of all the equities in the market over the safe rate of return. This is a measure of the risk premium on the equity market as a whole.

The risk premium on a particular company is the product of its beta and this average equity risk premium. The risk premium is the amount by which the required return differs from the safe rate. A company with a beta of 1 behaves like an average equity and the CAPM equation (3.1) implies it will have the average equity risk premium.

Beta is a measure of risk attached to a particular investment or company. If an investment in the equity of a company is not risky, the beta will be 0. In such circumstances equation (3.2) would show that equity investors should not expect a return from investment in that company to be different from that available on a safe asset. A safe asset is one where there is no uncertainty about the rate of return. (For investments over short horizons, yields on Treasury Bills or on government bonds are reasonable proxies for the safe rate). However, most companies' investments are more risky than safe investments with the result that those companies have a beta of more than zero.

Beta is an indicator of the extent to which the returns on the equity in a company fluctuate in line with general equity returns in the stock market as a whole. Such fluctuations are

[1]Intertemporal models will be considered in section 3.4.

costly to investors who dislike risk because they are difficult to avoid. If the returns on the equity of a particular company typically follow the returns on a portfolio of all equities then the risk of investing in that company is impossible to avoid by diversification. Beta measures the extent to which the returns on a specific company typically follow the returns on a diversified portfolio. Thus it is a good measure of the risk of investing in a company to an investor who holds a diversified portfolio of assets.

Beta is usually calculated by reference to recent movements in the company's share price; more specifically, it is assessed by estimating their covariance (or comovement) with returns on a diversified portfolio of stocks. If a company's share price usually moved in line with the market (that is by the same percentage) its beta would be 1.

The CAPM is widely used in the calculation of the required return on equity. Almost all regulators of utilities companies estimate acceptable rates of profit by reference to the CAPM. Use of the CAPM to estimate the required rate of return on the equity of a company is the usual procedure in large investment banks and securities houses. For example, Merrill Lynch, one of the world's largest investment houses, in its recent publication "The Cost of Capital Guide" uses the CAPM to estimate the required rate of return on the equity of companies throughout Europe. The London Business School share price service has for many years provided the inputs needed to use the CAPM to estimate the cost of equity. This service is sold at commercial rates. In a recent survey by Graham and Harvey (2001), three out of four chief finance officers said that they use the CAPM to calculate the cost of capital.

3.2.1. The Theoretical Basis of the CAPM

The CAPM predicts that the rate of return on a risky asset is a linear combination of just two components: the risk-free rate, the equity risk premium, with the weights given by the asset's beta. This simplicity is very attractive and largely explains the popularity of the CAPM for practitioners. The simplicity has a price, however: quite strong assumptions must be made. The exact assumptions made depend on the way in which the CAPM is derived. There are two standard routes: through the mathematics of the efficient portfolio frontier; and through a model of consumption. The latter will be reviewed in sections 3.3

and 3.4.

Two types of assumption underlie the portfolio frontier derivation of the CAPM. The first type can be shared by the CAPM and other asset pricing models.

ASSUMPTION 1: **Risk Aversion and Competitive Equilibrium**

- *Investors are risk averse.*
- *Markets for risky assets are in perfectly competitive equilibrium i.e.,*
  - *there are no transactions costs, taxes, constraints on short-selling, or other market frictions;*
  - *assets are infinitely divisible;*
  - *there is perfect competition (no individual investor can affect asset returns);*
  - *unlimited borrowing and lending is permitted*
  - *investors have identical beliefs about asset returns.*

The second type of assumption is particular to the static CAPM, and gives the CAPM its special features as expressed in equation (3.1).

ASSUMPTION 2: **Fund Separation**

The distribution of asset returns belongs to the class of separating distributions identified by Ross (1978); this class includes the multivariate normal distribution.

The portfolio frontier assumption ensures two-fund separation—given any portfolio of assets, there exists a portfolio of two mutual funds that investors prefer at least as much as the original portfolio. This gives immediately the characteristic form of the CAPM equation as a linear combination of returns on two portfolios.[2]

[2]To be precise: (i) two-fund separation means that all investors hold a linear combination of two mutual funds; (ii) by the definition of two-fund separation, the mutual funds are frontier portfolios (i.e., belong to

What happens if these assumptions do not hold? Clearly, the assumption of risk aversion for investors is crucial: without this, the trade-off between risk and return expressed by the CAPM does not hold. Certain deviations from a perfectly competitive market can be tolerated; for example, a version of the CAPM continues to hold even when borrowing is constrained—see Black (1972). If two-fund separation does not hold, then the resulting asset pricing model will be neither linear nor have a single factor.

3.2.2. The Empirical Support for the CAPM

If expected returns and betas were known and the 'market portfolio' were clearly identifiable, then an empirical test of the CAPM would be straightforward: simply plot the return and beta data against each other and test for a linear relationship. Unfortunately, neither expected returns, betas nor the market portfolio are known; in order to perform empirical tests, each must be estimated. This raises three problems:

- The CAPM implies a relationship concerning *ex ante* risk premia and betas, which are not directly observable.
- The CAPM as expressed in equation (3.1) is a single-period (although we will consider intertemporal versions in section 3.4. The data used to test the model are typically time series, as well as cross-sectional. Hence it is typically necessary to add an assumption concerning the time-series behaviour of returns. The simplest is to suppose that the CAPM holds period by period i.e., that returns are independently and identically distributed over time. It is unlikely, however, that risk premia and betas on individual assets are constant over time (a problem reviewed further in section 3.4).
- Many assets are not marketable; but, in principle, the CAPM requires that returns on the market portfolio, which includes all possible assets, be known.

the set of portfolios that achieve a given level of return with minimum variance); (iii) since the set of frontier portfolios is convex, all investors therefore hold a frontier portfolio; (iv) in equilibrium, the market portfolio is a convex combination of all investors' portfolios, and therefore is a frontier portfolio; (v) the mathematics of the portfolio frontier then places a linear restriction on expected asset returns in equilibrium that is shown in equation (3.1).

The standard solutions to each of these problems is to:

- Assume rational expectations, so that there is no systematic difference between *ex ante* expectations and *ex post* realizations; hence the former can be estimated by the former.
- Distinguish between *conditional* and *unconditional* versions of the CAPM. Even when risk premia and betas conditional on information sets available to investors over time are not constant, they can be constant conditional on a coarser information set.
- Proxy the market portfolio using a a major portfolio (such as a time-series of monthly rates of return on common stocks listed in the New York Stock Exchange (NYSE)), assuming that (i) the disturbance terms from regressing the asset returns on the return on the proxy market portfolio are uncorrelated with the true market portfolio; and that (ii) the proxy portfolio has unit beta. The sensitivity of tests to the proxy used can be examined.

A general empirical version of the CAPM is

$$Z_{jt} = \alpha_{jt} + \beta_{jt} Z_{mt} + \epsilon_{jt}$$

for $j = 1, \ldots, N$ and $t = 0, 1, \ldots, T$. Z_{jt} is the realized excess return (i.e., realized return minus the risk-free rate of return) for asset j at time t. Z_{mt} is the time-t market portfolio excess return. β_{jt} is the beta of asset j at time t; α_{jt} is the asset return intercept. ϵ_{jt} is the disturbance for asset j at time t; it is assumed to be uncorrelated with the excess market return. There are three ways that this model has been brought to the data:

1. A cross-section regression model of average (e.g., monthly) excess rates of return against estimated betas. That is,

$$\bar{Z}_j = \alpha + b\beta_j + \epsilon_j$$

where $\bar{Z}_j$ is the average excess rate of return of asset j. In this regression, β_j is treated as a fixed independent variable (estimated in some other procedure). If the CAPM is the correct model, then the regression coefficient b is the excess return on

the market portfolio. For example, Blume and Friend (1973) use this version to test the hypotheses that α equals zero and $b > 0$; this can be taken as a weak test of the CAPM's predictions. They find that both α and b were strictly positive; this finding is statistically significant.

2. A series of monthly cross-sectional regressions involving the realized excess rates of return:

$$Z_{jt} = \alpha_t + b_t \beta_j + \epsilon_{jt}$$

for all $j = 1, \ldots, N$ and $t = 0, 1, \ldots$. Here again β_j is treated as a fixed independent variable; if the CAPM is the correct model, then the regression coefficient b_t is the excess return on the market portfolio at time t. For example, Fama and Macbeth (1973) test the weak CAPM predictions that

$$\hat{\alpha} \equiv \sum_{t=0}^{T} \frac{\alpha_t}{T} = 0, \qquad \hat{b} \equiv \sum_{t=0}^{T} \frac{b_t}{T} > 0$$

when return distributions are stationary over time. Similar to the findings of Blume and Friend (1973), Fama and Macbeth find that both $\hat{a}$ and $\hat{b}$ were significantly strictly positive.

3. A series of time-series regressions for each asset/portfolio in the sample:

$$Z_{jt} = \alpha_j + \beta_j Z_{mt} + \epsilon_t$$

for all $j = 1, 2, \ldots, N$ and $t = 0, 1, \ldots, T$. In this model, β_j is a parameter that is to be estimated, while Z_{mt}, the excess return on the market portfolio at time t, is the independent variable. Note that the α_j and β_j are assumed to be constant over time. For example, Black, Jensen, and Scholes (1972) test the weak prediction that

$$\sum_{j=1}^{N} \frac{\alpha_j}{N(1 - \beta_j)} = 0.$$

If the Sharpe/Lintner CAPM is the correct model, this condition should not be rejected; if the Black (1972) version of the CAPM without a risk-free asset holds, then the

condition should be rejected. In Black, Jensen, and Scholes's test, the condition is rejected.

All three early tests of the traditional CAPM provided, therefore, weak support for the Black (1972) version of the CAPM without a risk-free asset. Two problems arose, however. The first related to the considerable econometric difficulties (as distinct from the conceptual difficulties discussed above) in testing the model. These called for techniques more sophisticated than standard ordinary least squares (OLS) to be used.[3] The second problem appeared in the late 1970s with a series of papers discovering 'anomalies'—firm characteristics that provide explanatory power for the cross section of asset mean returns beyond the beta of the CAPM. Many such anomalies have now been 'discovered'; the following is an incomplete list:

- *Small Firm Effect*: Smaller firms have higher expected returns than predicted by the CAPM. See Banz (1981), Keim (1983) and Reinganum (1983).
- *Value Effect*: Firms with low ratios of book value to market value have higher expected returns. See Chan and Chen (1991).
- *Neglected Firm Effect*: Firms with low institutional holdings have higher returns. See Arbel and Strebel (1983).
- *Overreaction*: Stocks which are down in one time period tend to rebound in the next (and vice versa). See De Bondt and Thaler (1985).
- *January Effect*: The return in January is consistently larger (by up to 8%) than returns for all other months. See Keim (1983) and Roll (1983).
- *Monday Effect*: The return from Friday close to Monday close is negative. See French (1980) and Gibbons and Hess (1981).

[3]Returns on financial assets may exhibit conditional heteroskedasticity, serial correlation and non-normal distributions. Such features call for the use of Hansen (1982)'s generalized method moments (GMM). In addition, maximum likelihood methods should be employed to test whether or not the market proxy portfolio is on the portfolio frontier, due to the nonlinear constraint imposed by the CAPM on the return-generating process.

The ever-growing anomaly literature presents a considerable challenge to the CAPM. One response has been to go on the offensive on three fronts. The first is to argue that there is little theoretical motivation for the inclusion of particular firm characteristics in an asset pricing model; see for example Brennan and Xia (2001). The second is to point out weaknesses in the methodology of the anomalies literature, such as neglect of the problems of survivor bias and data mining; see Kothari, Shanken, and Sloan (1995), and Campbell, Lo, and MacKinlay (1997). The third attack is the last line of defence for the CAPM: anomalies simply show that tests use poor proxies for the 'true' market portfolio. This harks back to the famous Roll (1977) critique of the CAPM—tests of the CAPM really only reject the mean-variance efficiency of the proxy; the model might not be rejected if the return on the true market portfolio were used.

A second response to the empirical questioning of the CAPM has been to move away from the linear, stationary and single factor features of the model. The alternatives are reviewed in turn in the next three sections.

3.3. Nonlinear Models

In section 3.2, we concentrated on a derivation of the CAPM that relied on fund separation and the properties of the efficient portfolio frontier. An alternative approach to the CAPM starts from a model of consumption. A consumption model analyses explicitly an individual's problem of maximizing the present discounted value of utility by choosing intertemporal consumption and investment in risky assets. The outcome of the analysis is an equation relating the ratio of marginal utilities across two periods to the asset price. This equation (known as the Euler equation) is, basically, a 'no arbitrage' condition—it requires that the individual cannot increase total utility by shifting wealth across the two periods. The ratio of marginal utilities measures how much total utility changes when a unit of consumption is shifted between periods; the asset price and expected return shows how much extra income (and hence consumption) can be generated by increasing investment by one unit. The condition is written

$$1 \;=\; \mathbb{E}_t \left[m_{t+1}(1 + R_{t+1}) \right] \tag{3.1}$$

where R_{t+1} is the random return of the asset at time $t+1$. m_{t+1} is known as the *stochastic discount factor*, or SDF; it is the discounted ratio of marginal utilities of consumption (in consumption-based models of asset pricing). Expectations at time t about time $t+1$ outcomes are denoted by the operator $\mathbb{E}_t$. See Cochrane (2001).

For empirical testing, it is more convenient to write equation (3.1) as a set of moment conditions:

$$\mathbb{E}[m(\mathbf{1} + \mathbf{R}) - \mathbf{1}] = 0 \tag{3.2}$$

where $\mathbf{R}$ is the vector of asset returns and $\mathbf{1}$ is a vector of ones. Three cases can be distinguished:

- There is a single factor and the SDF is a linear function of that factor i.e., $m = \beta f$. When the single factor is the market portfolio, this case generates the CAPM.
- There are several factors and the SDF is a linear function of the factors: $m = \sum_k \beta_k f_k$. This case generates a linear multifactor model, discussed further in section 3.5.
- There are any number of factors and the SDF is a nonlinear function of the factors. This case generates a nonlinear factor model.

This classification emphasizes the relationship between linear and nonlinear asset pricing models and allows an empirical test to distinguish between the two possibilities. To test for linearity, the SDF can be written as $m = \sum_k \beta_k f_k$; the parameters $\beta = (\beta_1, \ldots, \beta_K)$ can then be estimated by GMM using the moment conditions in equation (3.2). See Ferson (1995), Cochrane (1996) and Campbell, Lo, and MacKinlay (1997). This is the so-called SDF method of testing a linear asset pricing model; it can be contrasted to the more familiar beta methods that are discussed in section 3.2.[4]

[4]There is a methodological issue concerning the efficiency of the SDF method for non-linear models relative to the efficiency of the classical beta method for linear models. If the SDF method is relatively inefficient, then (as far as estimation efficiency is concerned) it is better to use the beta method and a linear pricing model. Kan and Zhou (1999) claim that the beta method is more efficient. Jagannathan and Wang (2001) dispute this conclusion, arguing that Kan and Zhou make inappropriate comparisons. They find that the SDF is asymptotically as efficient and has the same power as the beta method. The debate continues in Kan and Zhou (2001). See Campbell, Lo, and MacKinlay (1997), chapter 12 for a review of other approaches to estimating nonlinear models.

In this section, we put to one side the question of which factors should be used; this is dealt with in section 3.5. Here, we focus on the theoretical and empirical reasons for using a linear or nonlinear asset pricing model. It may seem intuitive that a nonlinear model will provide a better approach to asset pricing. There are three observations that qualify this intuition.

Any nonlinear model can be approximated by a linear model, in a variety of ways. The simplest way would be to take a linear approximation to a nonlinear SDF. Suppose that the 'true' SDF is a nonlinear function of a single factor f: $m = g(f)$ where $g(\cdot)$ is (for the sake of argument) a continuously differentiable nonlinear function. Consider the Taylor expansion for $m_{t+1} = g(f_{t+1})$ around the point f_t:

$$m_{t+1} = g(f_t) + (f_{t+1} - f_t)g'(f_t) + \left(\frac{(f_{t+1} - f_t)^2}{2}\right) g''(f_t) + \ldots$$

If the higher-order terms can be neglected, then this expansion becomes the approximation

$$\begin{aligned} m_{t+1} &\approx \alpha_t + \beta_t f_{t+1}, \\ \text{where} \qquad \alpha_t &= g(f_t) - f_t g'(f_t), \\ \beta_t &= g'(f_t). \end{aligned}$$

This approximation will be a good one only if the factor does not change too much i.e., $|f_{t+1} - f_t|$ is small, and if the higher-order derivatives of $g(\cdot)$ are not too large, evaluated at f_t. As Cochrane (2001) points out, the shorter the time interval, the less will be the variation in the factor. And even if this approximation is not a good one, others, derived by taking different expansion points (e.g., the conditional mean of the factor $\mathbb{E}[f_{t+1}]$), may prove satisfactory. In many cases, then, even if the 'true' model is nonlinear, a linear approximation will prove satisfactory.

In some cases (for example, where there is substantial variability in factors and/or the time interval between data observations is large), however, a linear approximation may not suffice. This leaves the challenge of estimating a nonlinear model. There has been a limited amount of work in this area.[5] The earliest work appears to be Kraus and Litzenberger (1976),

[5]The following is an incomplete list of work. One difficulty in reviewing nonlinear models is that the

(1983), which derive a three-moment asset pricing model by assuming that investors have a preference for positive return skewness in their portfolios. The SDF of this model is quadratic in the market return. More recent work includes Friend and Westerfield (1980), Scott and Horvath (1980), Sears and Wei (1985), Lim (1989), Leland (1999), Bansal and Viswanathan (1993), Bansal, Hsieh, and Viswanathan (1993), and Harvey and Siddique (2000). These papers typically find that nonlinear variables help to explain the cross-sectional variation of expected returns across assets, and that the effect is economically important. For example, Harvey and Siddique find a statistically significant risk premium on conditional skewness of 3.6% per year. Bansal and Viswanathan (1993) and Bansal, Hsieh, and Viswanathan (1993) analyse a class of nonlinear Arbitrage Pricing Theory (APT) models.[6] They find that nonlinearity of the SDF in the market return is significant in explaining stock returns.

Recent work has questioned some of these findings. Kan and Wang (2000) argue that a conditional CAPM (which allows for non-constant parameter estimates—see section 3.4) accounts for a panel of expected returns on equities better than a nonlinear APT model. (To be precise: using GMM estimation, they find that the Hansen and Jagannathan (1997) distance measure of pricing errors i.e., the maximum pricing error of the conditional CAPM is lower in all cases studied than that of a nonlinear APT model.) The issue that they highlight is that a carefully specified conditional CAPM—i.e., one in which appropriate time variability is allowed in the parameters of the linear model—will usually perform better than a nonlinear model. Those papers that conclude otherwise do so because they do not specify the conditional linear model correctly.

There is, in addition, a methodological problem in the estimation of nonlinear models: that of *overfitting*. An overfitted model fits the sample data "too well" i.e., it fits both the underlying deterministic components and the random errors. It will, therefore, perform very well within-sample (by definition); it is likely to perform far less well out-of-sample. Nonlinear estimation is particularly prone to the temptation of overfitting, since it is not clear how many degrees of freedom are given to the researcher. This problem is compounded by the fact that there is no one method that can test for the problem of overfitting.

set of nonlinear models is very large, being the complement of the set of linear models. We focus on asset pricing papers, and so do not include papers on nonlinearity in general macroeconomic data, for example.

[6] See section 3.5 for further discussion of the APT.

Finally, recent work has shown that in many cases, non-linear models are in fact 'higher moment' models that are linear in the relevant parameters. Satcehll and Hwang (1999) show that the CAPM can be extended to take account explicitly of measures of skewness and kurtosis. The resulting linear model can be estimated using the distribution-free GMM.

This section has provided a brief review of nonlinear asset pricing models. In doing so, it has pointed to the close connection, for empirical testing, between nonlinear and conditional models.

3.4. Conditional Models

The standard i.e., unconditional version of the CAPM asserts that the parameters in the relationship between an asset's expected excess return and the expected excess return on the market portfolio are constant; see equation (3.2). This is a product of the static framework in which the CAPM was originally derived. The development over the last two decades of consumption based models have highlighted the intertemporal factors that influence asset pricing. Recent research has documented mounting evidence of time-varying betas and time-varying risk premia; see, for example, Fama and French (1988) and Lettau and Ludvigson (2001a). Consequently, it has become common to allow for time-varying parameters in the CAPM:

$$\mathbb{E}_t[Z_{j,t+1}] = \beta_t \mathbb{E}_t[Z_{m,t+1}]$$

where $Z_{j,t+1}$ is the time $t+1$ excess return on asset j, $Z_{m,t+1}$ is the time $t+1$ excess return on the market portfolio, β_t is a time-varying factor loading, and $\mathbb{E}_t$ denotes expectations conditional on information held at time t. See Cochrane (2001), chapter 8 for further derivations. Numerous papers have adopted this approach: see Shanken (1990), Cochrane (1996), Jagannathan and Wang (1996), Ferson and Harvey (1999), Campbell and Cochrane (2000), Ferson and Siegel (2001), and Lettau and Ludvigson (2001b), amongst many others.

The general finding is that conditional models that allow for time-varying parameters in the asset pricing equation can perform substantially better than unconditional models. As Lettau and Ludvigson (2001b) comment, conditioning improves the fit of (in their case) the

CCAPM because

> some stocks are more highly correlated with consumption growth in bad times, when risk or risk aversion is high, than they are in good times, when risk or risk aversion is low. This conditionality on risk premia is missed by unconditional models because they assume that those premia are constant over time. [p. 1241]

Lettau and Ludvigson find that conditioning a CCAPM on the log consumption-wealth ratio gives a model that performs (i) better than an unconditional specification; (ii) about as well as the Fama and French three-factor model in explaining the cross section of average stock returns. Moreover, as Cochrane (2001) notes, conditioning (once the conditioning instruments are chosen) is simple to implement—it is just a matter of scaling unconditional factors by the chosen instruments and then proceeding in a standard way.

This route certainly provides a lifeline to the CCAPM, troubled as it has been by its inability to explain observed values of either the risk-free rate or the equity premium. In a sense, it is not surprising that conditioning on extra information improves the performance of the model. With a cunning choice of conditioning variables, the sample data can be matched very well to a (conditionally) linear CAPM. This is just the problem of overfitting encountered in section 3.3. The focus then shifts to the economic motivation for the variable brought into the model for the conditioning. (This is the explanation of the enduring appeal of the CAPM and the CCAPM, even in the face of empirical challenges to these models—both have a sound economic story to them.[7]) There is no accepted method to assess the extent to which conditioning factors have been chosen to fit data. Nor is there a consensus about what constitutes a compelling economic story for choosing a particular conditioning factor.

The difficulty is particularly acute in conditional CAP models. Conditioning is intended to reflect the information that investors use when making their consumption and investment decisions. This information is, of course, unobservable to the researcher. Hence there is no test available to assess whether the correct conditioning instruments have been used. When

[7]Cochrane (2001) notes that the CAPM is still taught despite the challenges. He notes that "[i]t seems that 'it takes a model to beat a model'" (p. 302). Much the same sentiment is expressed by William Sharpe in a 1998 interview that can be found at `http://www.stanford.edu/~wfsharpe/art/djam/djam.htm`.

the standard CAPM fails to explain stock returns, the defence implied by the Roll critique is that the market portfolio has not been specified correctly. When a conditional CAPM fails, a defence is that incorrect conditioning instruments have been used. There is no way to test whether this defence is correct or not.

These criticisms are expressed in Brennan and Xia (2002), who argue that the predictive power of Lettau and Ludvigson's conditioning variable (which they call '*cay*') comes mainly from a 'look-ahead' bias that is related to overfitting. Indeed, they show that the out-of sample predictive power of *cay* is negligible except in the first half of the sample period; and that a variable based on calendar time (which Brennan and Xia call '*tay*') is able to forecast stock returns as well as *cay*. See Lettau and Ludvigson (2002) for a response.

In summary: conditional CAPM i.e., models in which the parameters of the CAPM are time-varying offer some promise for improving the performance of the CAPM. But, like nonlinear models, the conditional models are susceptible to the charge of overfitting. Despite the large amount of work in the area, the methodology is some way from being agreed and testable.

3.5. MultiFactor Models

The anomalies described in section 3.2 have lead many to question the validity of the CAPM. An obvious response to the anomalies is to include in the asset pricing model additional factors related to the anomalies. For example, if smaller firms have higher expected returns than predicted by the CAPM, then including a variable related to firm size should help to improve the explanatory power of the model. (The obvious dangers in this approach are discussed below.) This somewhat *ad hoc* approach is given theoretical grounding by the Arbitrage Pricing Theory, developed by Ross (1976), which approaches asset pricing from a different direction to the standard CAPM argument. In this section, we first review the APT to assess the theoretical basis for multifactor asset pricing models. We then discuss the empirical evidence for multifactor models, concentrating on the seminal papers by Fama and French.

The CAPM starts from an explicit model of investor behaviour (mean-variance preferences, first analysed by Markovitz (1952)). In contrast, the APT of Ross (1976) starts with a more primitive assumption: that there should be no *arbitrage opportunity* in an economy. An arbitrage opportunity exists (roughly speaking) if there is no costless portfolio (in an economy with a large number of assets) the expected return of which is bounded below away from zero and the variance of which is negligible. A less technical, but more intuitive phrase is that there should be *no free lunch*. The economic assumption is that, in an economy in which agents are doing the best that they can to maximize their utility, any free lunch would quickly be eaten. In addition, the APT assumes that the payoff of a risky asset j is generated by K factors, $(f_1, f_2, \ldots, f_K)$ in a linear way: that is,

$$X_j = \alpha_j + \sum_{k=1}^{K} \beta_{jk} X_k + \eta_j. \tag{3.3}$$

In this equation, X_j is the payoff from asset j, R_k is the payoff from factor k, α_j and β_{jk} are asset- and factor-specific constants, and η_j is a random variable (idiosyncratic risk) with zero mean and covariance with the factor returns.

The APT uses the two assumptions—no arbitrage opportunities and the linear factor equation—to derive a prediction about *expected* rates of return in risky assets. Note that equation (3.3) refers to *realized* returns; it is a statistical characterization and has no economic content. If there were no idiosyncratic risk, so that $\eta_j = 0$ for all assets j, then it would state that

$$X_j = \alpha_j + \sum_{k=1}^{K} \beta_{jk} X_k.$$

When there are no arbitrage opportunities, this equation implies that the price of asset j can depend only on the prices of the K factors. In turn, this implies that the expected return on asset j is a linear function of the expected returns on the factors.

In practice of course, there is idiosyncratic risk i.e., $\eta_j \neq 0$. Nevertheless, the linear relationship between an asset's expected return and the expected returns on the factors,

which the linear factor equation (3.3) and no arbitrage imply holds exactly when $\eta_j = 0$, may hold *approximately*. The smaller is the idiosyncratic risk, the better the approximation. In fact, the APT shows that, when there is sufficiently large number of risky assets,

$$\left| (\mathbb{E}[R_j] - R_f) - \sum_{k=1}^{K} \beta_{jk} \left(\mathbb{E}[R_k] - R_f \right) \right| \leq \epsilon$$

where R_f is the return on the risk-free asset and $\epsilon > 0$ is some small number. In words: a linear relationship between the expected rates of returns on risky assets and the factors holds approximately for most of the assets when the economy is large. See Huang and Litzenberger (1998) for a more detailed derivation and explanation of the APT.

3.5.2. Consumption and Intertemporal CAPMs

There are two other types of multifactor model, both based on the CAPM: the consumption CAPM (or CCAPM, also discussed in sections 3.3 and 3.4), first developed by Breeden (1979) and the intertemporal CAPM (or ICAPM), first developed by Merton (1973).

In the APT, factors are any assets or portfolios that account for systematic correlation among asset returns; the factors arise from an arbitrage-based argument. In the ICAPM, it is assumed that there exists a limited number of 'state variables' (e.g., technology, employment income, the weather) which are correlated with assets' rates of return. In the CCAPM, the most important factor is aggregate consumption (or anything correlated with it). The ICAPM and CCAPM are closely related, both being based on equilibrium conditions and the utility maximization problems of investors; the CCAPM can be viewed as a special case of the ICAPM.

Despite the different origins of the APT and ICAP/CCAP models, they can all be collapsed into single beta (i.e., CAPM-like) representations when there are no arbitrage opportunities. No arbitrage implies the existence of a stochastic discount factor m_t (see section 3.3), so that

$$\mathbb{E}_t \left[m_{t+1}(1 + R_{t+1}) \right] = 1 \tag{3.4}$$

Model	Risk Factors	Portfolio p
APT	Common factors that account for systematic correlation among asset returns	Efficient combination of factor replicating portfolios
ICAPM	Growth rates of state variables including aggregate wealth	Efficient combination of market and state variable hedge portfolios
CCAPM	Aggregate intertemporal marginal utility function	Portfolio maximally correlated with marginal utility function

Table 3.1: Summary of Multifactor Models

where R_t is the return on the asset at time t. Further, the SDF can be written as a linear function

$$m_t = \gamma_t + \delta_t R_{\Phi,t+1}$$

where $R_{\Phi,t+1}$ is the time $t+1$ return on some portfolio Φ that is formed to replicate the characteristics of the SDF. γ_t and δ_t are (time-varying) parameters in the linear relationship. See Harrison and Kreps (1979) and Hansen and Jagannathan (1991). Straightforward manipulation of these two equations gives the linear equation:

$$\mathbb{E}_t[R_{t+1}] = R_f + \beta_{\Phi,t+1}\left(\mathbb{E}_t[R_{\Phi,t+1}] - R_f\right)$$

where $\beta_{Phi,t+1}$, like the CAPM beta, depends on the covariance between the returns of the asset and the portfolio Φ; R_f is the risk-free return.

The asset pricing models can, therefore, all be represented in the same way—as single beta models. The models differ in their specification of (i) the SDF, and hence the sources of uncertainty that drive the SDF; and (ii) the portfolio Φ. Table 3.1, taken from Lehmann (1992), summarizes the differences.

3.5.3. Summary of Empirical Tests of Multifactor Models

In this section, we provide a brief review of multifactor models, before concentrating on the most widely-cited and influential work by Fama and French (1992) in the next section. In the light of the previous discussion, we do not distinguish between APT, CCAP and ICAP models.[8]

The first empirical test of the APT, Gehr (1975), was in fact published before the original APT article of Ross (1976). The first comprehensive test is Roll and Ross (1980), who find evidence for three and perhaps four factors for generating process of returns. Chen (1983) estimates an APT model on daily return data, finding that the APT model performs well relative to the CAPM. He also finds that variables such as own variance and firm size do not contribute additional explanatory power to that of the factor loadings. Chen, Roll, and Ross (1986) is one of the first papers to employ macroeconomic variables as APT factors, rather than choosing factors by some statistical analysis of security returns. They find that risk from the spread between long and short interest rates, expected and unexpected inflation, industrial production, and the spread between high and low grade bonds are all significantly priced. (They also have results on macroeconomic factors that are not priced separately.) Shanken (1990), Ferson and Schadt (1996), Jagannathan and Wang (1996), and Cochrane (1996) all extend the CAPM by scaling the market factor with variables such as the dividend-price ratio or returns on physical investment. These models are therefore unconditional multifactor models. All of the papers report reduced pricing errors, relative to the standard CAPM, by introducing additional factors.

Several empirical difficulties have been encountered in the implementation of the multifactor models. The APT is silent about what the factors are: it simply assumes that they exist and that a linear combination of their realized returns determines the realized return of risky assets. There has been a lengthy debate about (i) how to determine the number of factors (see Connor and Korajczyk (1993)); (ii) whether the 'true' number of factors is likely to be large (see Dhrymes, Friend, and Gultekin (1984)); and (iii) whether it matters (see Lehmann and Modest (1987)). The APT gives an approximate relationship; for any

[8]For a very comprehensive reading list on multifactor models, see `http://www.kellogg.nwu.edu/faculty/korajczy/htm/aptlist.htm`.

given asset, the deviation of its expected rate of return from the APT relation might be very large. Most APT applications assume that the pricing errors are negligible and use the pricing equation as if it were a strict equality. The empirical weakness in the ICAPM is that the set of variables that may serve as proxies for changes in investment opportunities is too broad. Essentially, any variable that forecasts future returns would be priced in the ICAPM. On the other hand, the CCAPM provides a factor model with aggregate consumption as the only possible factor. An mentioned elsewhere in this report, analyses using only aggregate consumption as a factor have proved disappointing empirically.

3.5.4. The Fama and French MultiFactor Model

Fama and French (hereafter FF) (1992), (1996) attempt to resolve two of the key anomalies that have plagued the empirical CAPM, by including two additional factors into the asset pricing equation. In doing so, FF provide evidence that this also eliminates any marginal impact of a wide range of further anomalies (such as the dependence of a firm's equity returns on its past sales growth).

The empirical basis for the FF model is derived from time series regression equations, for the excess return on asset or portfolio i of the form (as in equation (2) of FF (1992)):

$$R_i - R_f = \alpha_i + \beta_i(R_m - R_f) + s_i SMB + h_i HML + \epsilon_i$$

where SMB is the difference between the returns on two portfolios, one of small, and one of large stocks; and HML is the difference between the returns on two portfolios, one of high, one of low book-to-market ratios.

FF show that such equations have very high explanatory power (with a typical R-squared of around 0.95) for a range of portfolios, sorted according to a range of different prior characteristics. Earlier work (FF) showed explanatory power from the same factors for the cross-section of individual stock returns. The collective and individual significance of the factor loading parameters β_i, s_i and h_i is thus clear-cut. In most specifications, the restriction that the intercept a_i is zero cannot be rejected at standard levels of statistical significance.

On the basis of these time series regressions, FF infer a pricing equation (consistent with

the APT) where the risk premium on asset i is a linear combination of the risk premia on the three factors:

$$\mathbb{E}[R_i] - R_f = \beta_i(\mathbb{E}[R_m] - R_f) + s_i\mathbb{E}[SMB] + h_i\mathbb{E}[HML].$$

On the face of it, this pricing equation potentially offers economically significant differences from the expected returns implied by a simple CAPM. The key differences are:

- The historic risk premia associated with the two non-CAPM factors are large. In the FF sample (1964–1993) the mean annual excess returns were 4.9% on SMB, and 6.3% on HML. The latter is higher than the excess return of the market over safe assets within the same sample, of 5.9%.
- Estimated values of the β_i are generally very close to unity. This has two implications. First, in the absence of the other factors, the pricing equation would have little explanatory power for the cross-section of expected returns (consistent with a range of evidence, from FF and others, that the CAPM beta does not explain mean returns). Secondly, the risk-free rate essentially disappears from the equation.

As a relevant example, Giles and Butterworth (2002), on behalf of T-Mobile, find fairly similar results to FF in an implementation on UK data, albeit with a very limited role for firm size (on which more below). They find a similar mean excess return on the UK equivalent of HML to that of FF; and, on the basis of estimated impact factors for Vofafone, calculate that this implies an additional impact on the cost of capital for T-Mobile of up to two percentage points, over and above that from the standard CAPM beta.[9]

It is worth noting, however, before proceeding to a more detailed critique of FF, that even at face value, the additional factors in the FF pricing equation do not necessarily have a major impact on estimated expected returns for most regulated industries. The explanation for this is straightforward: while the average firm will have a β_i of unity (and, indeed, so will most non-average firms, in FF's results), the values for the average firm of the two

[9]In their results, the UK HML is close to orthogonal to the market excess return, so the simple addition of the FF-type effect is defensible; on US data, the market return and HML are negatively correlated, so the FF beta does not correspond to the simple CAPM beta.

other loading factors s_i and h_i will be zero, since, e.g., high book-to-market firms will have a positive value of h_i, but low book-to-market firms will have a negative value. Since such firms are only identified in relative terms, the average effect must be zero. The impact of the two additional factors will be large only for firms that are at extremes of the cross-section.

We note below, in fact, that the role of the additional factors may be more significant in an indirect way, by changing estimates of the β_i.

There are also quite important reasons to be sceptical of the basis for the FF pricing equation. While the estimated loading factors are, as noted, highly significant, much less attention has been paid to the statistical significance of the risk premia themselves. How confident we can be that, on average, the risks associated with these factors have a positive price? A range of authors (e.g., Campbell (2001b) and MacKinlay (1995)) have criticised FF for 'data-snooping'—inferring the existence of such factors from the features of a single, relatively short sample, that might not be representative of the true underlying model.

Just as in the case of the equity risk premium, the assumption that mean returns on the two factor portfolios are equal to the mean risk premia requires the assumption that expectational errors have cancelled out over the sample. It is noteworthy that in other work (Fama and French (2001)), FF give substantial credence to the possibility that, in the case of the equity risk premium itself, expectational errors have not cancelled out in the post-war period, thus biasing upwards the estimate based on realized returns.

In the case of the two additional FF factors, the basis for assuming positive risk premia appears distinctly more fragile, for a number of reasons:

- FF themselves acknowledge that the theoretical basis for their factors is, at best, patchy. While they can, to a limited extent, be rationalized *ex post*, there is no clear theory that posits a positive premium on the factor portfolios. Indeed, if anything, there is one simple theory that posits a premium of precisely zero: namely, that, in a CAPM world, positive excess returns on the factor portfolios should represent an unexploited arbitrage opportunity.

- Even within their own sample, the empirical evidence of significantly positive premia is not very strong. Table 3.2 shows that, on the basis of standard t-tests (as reported in

	$R_m - R_f$	SMB	HML
1964–1993	5.94 (1.96)	4.92 (1.72)	6.33 (2.60)
1993–2002	6.61 (0.87)	0.26 (0.05)	-0.43 (-0.06)
1964-2002	6.09 (2.16)	3.84 (1.56)	4.77 (1.93)

Table 3.2: Sample Arithmetic Means (t-statistics in parentheses)

FF (1996)), the sample mean of the excess return SMB is of only marginal significance; the statistical significance of the mean excess return on HML is called into question by the possibility of data mining.

- We have extended the two original series in FF (1996) out of sample, using reasonable proxies.[10] Table 3.2 shows that in the sample since 1993, these proxies have means that are insignificantly different from zero. In the full sample, this implies that the t-statistics for the null hypothesis of a zero mean are distinctly more marginal, even at classical significance levels.[11]

- In contrast, for comparison, table 3.2 shows that adding an additional 8 years of data somewhat increases the precision of the estimate of the equity risk premium, as it should do if the underlying true mean value is constant.

Figure 3.1 illustrates the fragility of the estimated premia by cumulating excess returns on the two factor portfolios. It shows, as noted by others (e.g., Dimson, Marsh, and Staunton (2001b) and Siegel (1998)), that the reliability of the small firm effect was already looking fairly suspect even by the time of FF's estimation work. The book-market effect survived for longer, but went into full-scale reverse during most of the 1990s.

It would be almost certainly be premature to conclude from this that available evidence for the positive risk premia is definitely spurious. FF note, in countering arguments of

[10] HML is proxied by the return on the Barra value index, less that on the Barra growth index (these partition the S&P 500, largely on the basis of book-to-market ratios). In the common sample (1974–1993) the proxy has a correlation of 0.87 with HML. It is rescaled in extrapolation. SMB is proxied by the return on the S&P SmallCap 600, less that on the S&P 500. Unfortunately no common sample is available. Data for 2002 are for August.

[11] Note also that the estimated mean geometric returns are even closer to zero, and even less significant—an element in the positive arithmetic mean is simply due to the variance of excess returns on the portfolios.

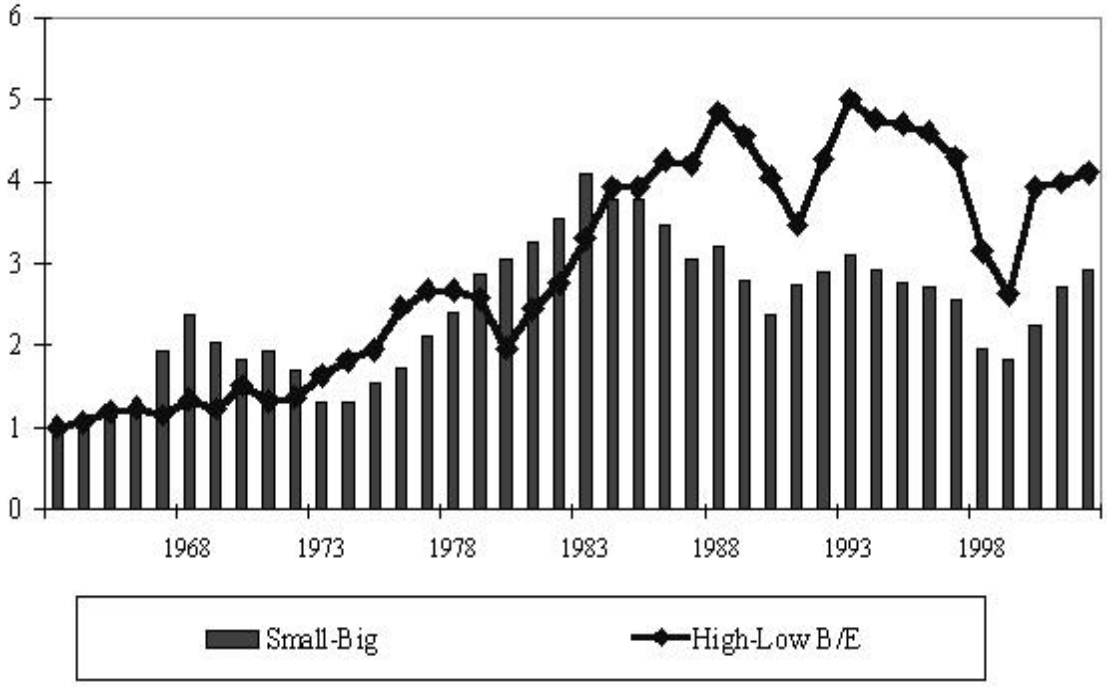


Figure 3.1: Cumulative Returns on Fama and French Factor Portfolios

data mining, that both risk premia appear to have been significant in earlier periods; and Dimson, Marsh, and Staunton show that they have also been present in the UK and other countries. But a conservative position is that the case for positive premia on these factors is 'not proven'.

This conclusion must be treated with caution, however, when dealing with FF-style pricing equations. On the face of it, if we assume that the expected premia on the two additional factors are zero (an assumption that does not appear inconsistent with history, nor with simple theory), then, despite the predictive power of these factors period-by-period, in expectation we are left with simply the first term in the FF pricing equation—which looks just like the CAPM.

However, deriving the CAPM in this indirect way will not have the same implications for the estimated expected cost of capital, since, as noted above, the estimated values of the β_i in the FF equations are almost all very close to unity. This is indeed an inescapable conclusion to draw from the research of FF and others, that finds that the CAPM beta has very weak explanatory power for the cross-section of returns. FF's results imply that, in contrast, standard estimated values of the CAPM β that differ significantly from unity may be proxying the impact of the omitted factors. Yet if these factors themselves have associated risk premia of zero, allowing the CAPM beta to vary significantly from unity may be throwing too much or too little weight on the role of market risk, for any given stock, since, on the basis of FF's work, the underlying model used to estimate beta is mis-specified.

At a minimum, therefore, the FF approach seems to provide additional ammunition in favour of approaches that tend to bias estimated CAPM betas towards unity, as in the Bayesian adjustments applied to LBS beta estimates.[12]

3.6. Conclusions

- The Capital Asset Pricing Model (CAPM) is (still) widely-used to estimate firms' costs of capital. There is considerable evidence of empirical shortcomings in the CAPM; but

[12]Taken further, it would point towards estimates of beta, derived from FF-style equations, that may be virtually indistinguishable from unity—implying that all that would matter for the estimated equity return would be the real aggregate stock return (i.e., the safe rate and the equity premium entering symmetrically).

its clear theoretical foundations and simplicity contribute to its popularity.

- The alternative of nonlinear models of asset pricing have not achieved such popularity. There are several reasons for this. The most important is the problem of 'data overfitting'. Nonlinear models are particularly prone to this temptation, since it is not clear how many degrees of freedom are given to the researcher. The problem is compounded by the absence of any one method that can test for the problem of overfitting. In addition, in many cases a nonlinear model can be approximated well by a linear model. Finally, recent research suggests that a carefully specified conditional CAPM—i.e., one in which appropriate time variability is allowed in the parameters of the linear model—will usually perform better than a nonlinear model.

- Conditional models, in which the parameters are time-varying, have been the focus of much recent work. As with nonlinear models, the problem of data overfitting is present; and there is no test to assess the extent of the problem. In addition, there is no test available to assess whether the correct conditioning instruments have been used. (Hence there is a current debate about whether a time trend variable is a better conditional variable than the consumption-wealth ratio.) Despite the large amount of work in the area, the methodology is some way from being agreed and testable.

- Multifactor models have also received considerable attention, particularly since the influential work of Fama and French (FF). The standard difficulty with multifactor models is the satisfactory identification of the factors. There has been, for example, a considerable debate about whether the small firm factor used by FF is relevant for other time periods and markets. The risk premia on the two additional factors used by FF are of marginal statistical significance in their study; when the sample period is extended to include later data, the premia are not statistically significant. The inclusion of the factors in the asset pricing model has the general effect of moving the CAPM beta (i.e., the factor loading on the usual market portfolio) towards 1.

- In summary: the empirical shortcomings of the CAPM are known. Alternative models to address this issue have their own shortcomings—weak theoretical foundations and empirical challenges. In our view, there is no one clear successor to the CAPM for practical cost of capital estimation.

4. Beta estimation

This section addresses practical issues in the estimation of betas. We consider the following:

- Optimal frequency of data with which to estimate betas.
- Choice of Estimation period.
- Choice of safe rate and definition of excess return.
- Choice of market index, with particular focus on the international mix of assets.
- Bayesian adjustments.
- Estimation of betas for companies with limited stock market data.

4.1. Data Frequency

4.1.1. Theory

It is common for betas to be estimated with weekly or monthly data on the return on the individual stock and on the market. The LBS Risk Measurement Service, for example, estimates betas using the most recent 60 months of returns on the stock and the market (taken to be the return on the FTSE All Share Index). For most stocks daily return data is available so there is an issue about whether use of weekly or monthly returns is inefficient. Are we throwing away valuable information by ignoring intra month or intra week movements in stock prices?

Under certain circumstances there is an unambiguous answer to this question. If we assume that stock price returns are serially uncorrelated and that the link between the

market return and the return on the individual stock is the same for all frequencies then using OLS estimates of beta based on the highest frequency data is optimal. The result is a straightforward implication of well known properties of OLS estimates.

Briefly the logic is this. Assume that the highest frequency data we have is, say, daily. The link between the daily excess returns (i.e., return over and above a safe rate) on stock i and the market is:

$$R_{it} = \alpha + \beta R_{mt} + e_{it} \tag{4.1}$$

where R_{it} is the log excess return on asset i at day t (i.e., it is the log return net of the logarithmic safe rate), R_{mt} is the log excess return on the market, α is a constant and β is the beta. e_{it} is the non-systematic component of the return to asset i at time t.

Assume that $R_{mt} = \mu + w_{mt}$ where μ is the mean return on the market and where w_{mt} is the random component of that return at time t. If the random components of returns (e and w) are independently and identically distributed (iid) then they have zero expectations, no correlations and constant variances:

$$\begin{aligned}
E(w_{mt}) &= E(e_{it}) = 0, \\
E[(w_{mt})^2] &= \sigma_m^2, \\
E[(e_{it})^2] &= \sigma_i^2, \\
E[e_{it}e_{it-j}] &= 0 \text{ for all } j \text{ except } j = 0, \\
E[w_{mt}w_{mt-j}] &= 0 \text{ for all } j \text{ except } j = 0.
\end{aligned}$$

The ordinary least squares (OLS) estimate of β, β_{OLS}, is:

$$\frac{\Sigma_{t=1}^{T}(R_{it} - R_i)(R_{mt} - R_m)}{\Sigma_{t=1}^{T}(R_{mt} - R_m)^2}$$

where R_i and R_m and the sample means of the daily excess log returns on asset i and the market. T is the total number of daily observations we have from which to estimate the Beta.

Assuming that the random components e and w are normally distributed, β_{OLS} is asymp-

totically normally distributed with a mean β and a variance of $(1/(T-2))(\sigma_i^2/\sigma_m^2)$ i.e.,

$$\beta_{OLS} \sim^a N[\beta, 1/(T-2)(\sigma_i^2/\sigma_m^2)]. \tag{4.2}$$

In finite samples the estimator follows an unconditional t distribution. For large samples, the t distribution will be very much like a normal distribution, but it has slightly thicker tails.

Suppose we form n period returns (eg we take daily data and form monthly returns by adding up 20 observations on daily log excess returns to give monthly log excess returns). This will generate T/n observations.

OLS estimation is still unbiased but will have a larger variance. Using n period aggregation we would have:

$$\beta_{OLS} \quad \sim N\left[\beta, \frac{1}{T/n-2}\frac{\sigma_i^2}{\sigma_m^2}\right].$$

This implies that for fairly large T, the standard error of the estimate will be roughly $\sqrt{n}$ times larger than if we use daily data.

To give a concrete example. Suppose:

$$\beta = 0.8; \quad \sigma_m = 0.175; \quad \sigma_i = 0.25.$$

Then the OLS estimate of β will, on average, be 0.8 and 95% of the time it will fall roughly in the interval of around 2 standard errors either side of 0.8. Assume we use tow years of data; either 500 daily observations or 25 monthly observations. The standard error of the estimate based on daily data (that is the standard deviation of the estimate) is

$$(1/\sqrt{500}) \times 0.25/0.175 = 0.064,$$

which means there is a 95% chance that the OLS estimate will be in the range 0.927 and 0.672.

With monthly data, assuming 20 trading days a month, the standard error is $1/\sqrt{(25-2)}0.25/0.175 =$

0.298, and the 95% confidence interval becomes 1.396 to 0.204.

Clearly there is a huge increase in accuracy with daily data.

It is essential to this argument that the assumptions of iid returns carries over to daily data. It is also essential that the relation between returns on the market on the individual stock is the same at 1 day as at a week or a month. Failure of these assumptions could make using a frequency greater than one day sensible. The issue then is how reasonable are these assumptions and if they fail what is the right way to handle that.

4.1.1.1. Failure of iid assumption There is evidence that at high frequencies returns may be correlated. In the US there is some evidence of positive correlation of daily returns from one day to the next (see in particular Section 2.8.1 of Campbell, Lo and MacKinlay (1997)). Serial correlation in weekly and monthly returns tend to be somewhat less significant, though not absent. It is important to note that this does not affect the unbiasedness of an OLS estimate based on daily data. But it does affect efficiency and, more important, will make estimated standard errors misleading.[1] Assuming there is much less correlation in returns from one observation to the next with monthly data, then the problem would be less with lower frequency data. But a better response thanusing monthly data is likely to be to use daily data and implement some form of serial correlation adjustment to standard errors rather than throw the baby out with bath water and go to monthly data.

Certainly if one were to use daily data one should calculate robust standard errors which are unbiased even if there is serial correlation. To this end one should use a version of the Hansen Hoderick or Newey and West (1987) procedure. Newey-West standard errors are consistent in the presence of serial correlation. The option to use them in place of unadjusted standard errors is now standard with most econometric software packages. What this does is adjust the reported statistics for the impact of serial correlation.

If heteroskedasticity appears as a problem with high frequency data—and it is *NOT* obvious it is more of a problem that with weekly or monthly data—White's (1980) het-

[1] In the presence of serially correlated errors OLS is unbiased and consistent; coefficient estimates are asymptotically normally distributed (assuming underlying processes driving residuals are normal) but inefficient.

eroskedasticity corrected standard errors can be computed[2]. The option to compute White's standard errors is also standard on most econometric software nowadays. Many packages now also produce estimates of standard errors that are robust to both heteroskedasticity and serial correlation of residuals[3]

4.1.1.2. Failure of the assumption that the relation between returns is the same at all horizons For infrequently traded stocks it may be some time before the impact of a general market movement shows up in the stock price. For large stocks it is very likely that any impact of general market conditions is reflected in transaction prices and quoted prices within the same day. Indeed one might find the opposite phenomenon with *very* highly traded stocks that the individual stock price moves in response to news ahead of it showing up in a general index of stock prices. This is a thick trading problem with using daily data as opposed to a thin trading problem.

For less frequently traded stocks where it may take more than a few hours for new information to be reflected in measured process a daily beta estimate is likely to be downward biased. For very heavily traded stocks the impact can go the other way leading to an overestimate of beta. Two procedures could be followed to handle the issue. First, moving to lower frequency (weekly/monthly) data will reduce the problem. Second, one could stick to daily data but include as the regressors in the estimating equation (4.1) the current, lagged and forward value of the return on the market. That is we estimate:

$$R_{it} = \alpha + \beta_1 R_{mt} + \beta_2 R_{mt-1} + \beta_3 R_{mt+1} + e_{it} \tag{4.3}$$

where R_{mt-1} is the return on day $t-1$ and R_{mt+1} is the return on day $t+1$.

The estimate of the CAPM beta is then $\beta_1 + \beta_2 + \beta_3$.

If the stock is very infrequently traded it could be important to add further lags.

The idea behind equation (4.3) is that for stocks that are either very thickly traded or very thinly traded we may miss some of the typical co-movement between the market and

[2]For details of the Newey-West and White's standard errors see, for example, Greene (1993).
[3]See, for example, Andrews (1991).

the individual stock by just focusing on co-movement on the same trading day. It may take a while for the price of the individual stock to adjust to the market in general if it is thinly traded—this is why we put in a lag of the market return in the equation to estimate overall co-movement (R_{mt-1}). For a very heavily traded stock (a thick market stock) it conceivably could take a while for the general market to catch up with news which is reflected in the stock's price almost instantly. This is why we include a lead of the market in the equation (R_{mt+1}).

By putting in leads and lags we may be able to preserve the efficiency advantages of using daily data. But if the thin trading is very serious we might need more than 1 daily lag. It would be less usual for a thick traded stock to require more than one daily lead of the market since we do not expect general market news to take more than 1 day to show up in the overall level of the markets. (More precisely, we do not expect it to take more than 1 day extra for news to show up in the price of the general market than in the price of a thickly traded stock.)

Of course putting in extra leads and lags adds somewhat to the process of calculating betas. More important, there is inevitably some uncertainty about how many leads and lags to include and a degree of arbitrariness about where to draw the line. All this is a disadvantage of using daily betas. Advantages of extra precision will need to be significant if this disadvantage, relative to using beats estimated on weekly or monthly data, is to be outweighed. The empirical evidence we describe below sheds some light on the scale of the extra precision using daily data can bring.

4.1.1.3. Data issues with daily data: bid-ask bounce Daily data on less frequently traded stock are also open to measurement problems with the individual stock return stemming from bid-ask bounce. This phenomenon can give spurious negative serial correlation in returns when a trade at the bid is followed by a trade at the ask. But this is only a problem if stock prices reflect the last trade at the actual price transacted (i.e., at the end of day bid or ask). In practice this is not likely to be a relevant issue in the UK. The London Stock exchange calculates closing prices for heavily traded stocks as the volume weighted average of all trades in the last 10 minutes of trade. If there has been no trade in the last 10 minutes the mid point of the best bid and ask is used as the closing price. So the bid-ask bounce

phenomenon is not, in practice, likely to be much of a problem.

In summary the problems with using daily data are likely to stem from infrequent or non-synchronous trading. Estimators which produce standard errors robust to serial correlation and heteroskedasticity (Newey West and White standard errors for example) and the use of leads and lags in the market index can, in principle, handle these issues. In general we would expect the gains in precision from having more observations, or the advantages in being able to rely upon more recent data, would outweigh the disadvantages of inefficiencies due to induced serial correlation, heteroskedsatisicty and other timing issues. One indication of the likely scale of the problems with using daily data is the existence of serial correlation and heteroskedasticty, so at least the existence of problems will be signaled. If there do seem to be significant signs of these problems with daily data and much less sign of them with weekly or monthly data this is an argument for looking at beta estimates based on the latter. Absent such problems using daily data has clear advantages.

One final advantage in using daily data is that no decision needs to be made over which day to use for measuring returns. For weekly or monthly data estimated betas can be sensitive to when in the week/month returns are measured

4.1.2. Empirical Evidence

Table 4.1.2 reports the results of estimating the beta of British telecom (BT) using daily, weekly, monthly and quarterly data. Here we use the FTSE all share index as the market portfolio. We use data from the 5 year period ending in August 2002. The table shows unadjusted standard errors and standard errors corrected for heteroskedasticity (White's standard errors) and corrected for both heteroskeadsticity and serial correlation.

There are several points to note in Table 4.1.2. First, the standard errors for the daily estimates are very much lower than with weekly, monthly or quarterly data. Standard errors from daily estimates are around one third the standard errors from estimates based on monthly data. This is close to what one would expect (i.e., a ratio of about the square root of $(60/1250) = 0.22$). Second the heteroskedasticty and serial correlation adjustments are significant—and somewhat more so for daily data than for weekly, monthly or quarterly

	No. of obs.	Beta	Unadjusted OLS standard error	Whites heteroskedasticity consistent standard error	Heteroskedasticity and serial correlation consistent standard error
Daily	1250	1.052	0.034	0.050	0.056
Weekly	260	0.960	0.074	0.087	0.088
Monthly	60	0.855	0.124	0.127	0.139
Quarterly	20	1.070	0.158	0.195	0.180

Table 4.1: Estimates of the beta of British Telecom—5 year regression window to August 2002

data. But the coefficient estimates are unbiased and are very tightly estimated for daily data so that any loss in efficiency from using OLS is likely to be small. (The two standard deviation interval for BT based on daily data is 0.94 to 1.16 using the most robust standard errors). Third, there is no obvious pattern to the central estimates—the estimate based on the quarterly data is higher than that based on weekly or monthly data and very close to the estimate based on daily data; this is in line with the consistency point made above.

Table 4.2 shows estimates of the beta for Vodafone based on daily, weekly and monthly data over a recent five year period[4]. This table reveals that the standard error of the beta estimated from daily date is around one quarter the beta estimated from monthly data and about one half that estimated from weekly data. This is not too far off the gain in efficiency that one would expect—based on a 20 day trading month we would expect the daily beta to have a standard error $1/\sqrt{20}$ of the estimate based on monthly data (ie. around 22% of the monthly level). The estimate based on daily data should have a standard error of the beta of about $1/\sqrt{5}$ (or 45%) of the weekly beta. Notice that these are unadjusted standards errors so the gain in efficiency is likely to be somewhat overstated.

The confidence interval on the monthly beta in Table 4.2 is so large that although the

[4]Reproduced from the paper "Issues in Beta Estimation for UK Mobile Operators", the Brattle Group, July 2002.

estimate seems very far from the estimate based on daily data (0.99 as against 1.65) it only represents about 1.35 unadjusted standard errors of the monthly estimate. If all the strong assumptions required to make the estimated standard errors reliable hold we can conclude the following from Table 4.2: based on the daily data we can be very sure that the true beta is in excess of 1.3 and very far above the central estimate based on the 60 monthly observations. Yet there is nothing particularly surprising about the monthly results; so great is the standard error of the monthly estimate that even though the best guess based just on that data is that beta is just under unity one could not rule out a beta of 1.65 at standard (5%) confidence intervals. This rather powerfully reveals the problem with monthly estimates—unless one uses data from well over 5 years ago the standard errors will generally be large.

Notice that the estimate of beta based on weekly data exceeds the estimate based on daily or monthly data. This suggests that one interpretation of the relatively high beta based on daily data—that it might be biased up by an outlier when Vodafone's share price and the market index both rose freakishly or fell freakishly on a single day—is not very convincing.

	Estimated beta	unadjusted standard error of estimate
Five years daily	1.65	0.13
Five years weekly	1.89	0.29
Five years monthly	0.99	0.49
Notes: Estimates made in July 2002.		

Table 4.2: Estimates of Vodafone Beta based on weekly, monthly and daily data

4.2. Choice of Estimation Period

4.2.1. General issues

There is a great deal of evidence that betas vary over time. This may reflect movements in gearing, which should have an impact upon equity betas, or changes in the underlying correlations between company and aggregate returns (i.e., variablility in asset betas). This

issue is directly relevant to the choice of estimation window. In the absence of an explicit method for handling time varying covariances and variances the best one can do in handling changing betas is to use as recent an estimation window as is consistent with estimates having low standard errors. The tradeoff between using an estimation window that gives low standard errors (which means having a large number of observations) and one which comes from a period where beta is likely to be close to its current value (which requires a short estimation window if there is time-variation) is much more favourable with daily data than with weekly or monthly data. This is one of the strongest reasons for using daily data. Six months of daily data will give about 120 observations—the equivalent of 10 years of monthly data. While a company beta, in the absence of some obvious change in the nature of the business, is unlikely to be dramatically different to the past over a 6 month horizon , it is often very different from its value 10 years ago.

It is important in this context to note that the gain in estimation accuracy—as measured by the fall in the standard error of the estimated beta—from having more observations becomes less as more observation are added. With daily data going from 250 observations to 500 observations (i.e., from a 1 year window to a 2 year window) will reduce the standard error by about 40 percent. Going from 2 years to 3 years will, other things equal, only reduce the standard error by about 22%; extending the estimation window another 1 year, from 3 to 4 years, reduces the standard error by a further 15%.

Use of an explicit technique to handle time-varying variances and co-variances is, in some ways, the ideal solution. Not only does this allow one to use data from periods when beta may have bee very different but it also allows one to project future changes in beta since a projected path for variances and co-variances can be derived from a model of the time series evolution of the moments of the data. (An example of the technique is Hall, Miles, and Taylor (1989). But a major drawback of the technique is that it is susceptible to over-fitting and can reveal apparent signs of time variation where none exist, especially if complicated models of time variation is used. At a more practical level, it involves use of techniques that are highly non-linear and not widely used amongst practitioners who estimate betas. So there would be a problem of getting a beta estimated with a time varying technique to be widely accepted as a standard estimate—this is partly because there are many different ways to model time variation (GARCH, EGARCH, GARCH in mean and many newer variants).

As with serial correlation and heteroskedasticity, there are at least some obvious tests of whether time variation is a problem—formal statistical tests of breaks in the process as well as less formal tests like observing the plot of how estimates of beta based on a fixed estimation window evolve (see below).

Our recommendation is that using between one year and two year periods with daily data will generally give low standard errors and that if the one year betas and two years betas are little different the time variation problem is unlikely to be significant. If those betas do look different one could estimate one year and six month betas and if these are little different use the one year beta.

4.2.2. Empirical Evidence

Figures 4.1–4.8 show the results of rolling regressions to estimate the beta of British telecom. We use daily, weekly, monthly and quarterly data from the period 1990 to August 2002. Each figure shows the evolution of the beta estimated on the latest 5 years of data. The estimates are updated at the end of each month. The firsts set of results use as the market index the FTSE all share index; the second set of graphs show the betas when we use a market index made up of 70% FTSE all share index and 30% the FTSE world stock index (converted into £ terms).

The graphs reveal apparent signs of major changes in beta when we use monthly or quarterly data. The monthly beta bade on the FTSE all share index moves from under 0.8 up to around 1.2 over the sample period. The daily and weekly betas, in contrast, look much more stable and rarely move much from unity. This illustrates one of the problems with using monthly data—a five year window only gives 60 observations and random fluctuations in estimated beta will arise as one observation is dropped and one added because the standard error of the estimate is large. This problem is smaller with weekly data and much lower with daily data. Notice that the standard errors with monthly and quarterly estimates are so large that despite the very large fluctuations in the estimated beta they hardly move outside the 2 standard error ranges estimated at the outset.

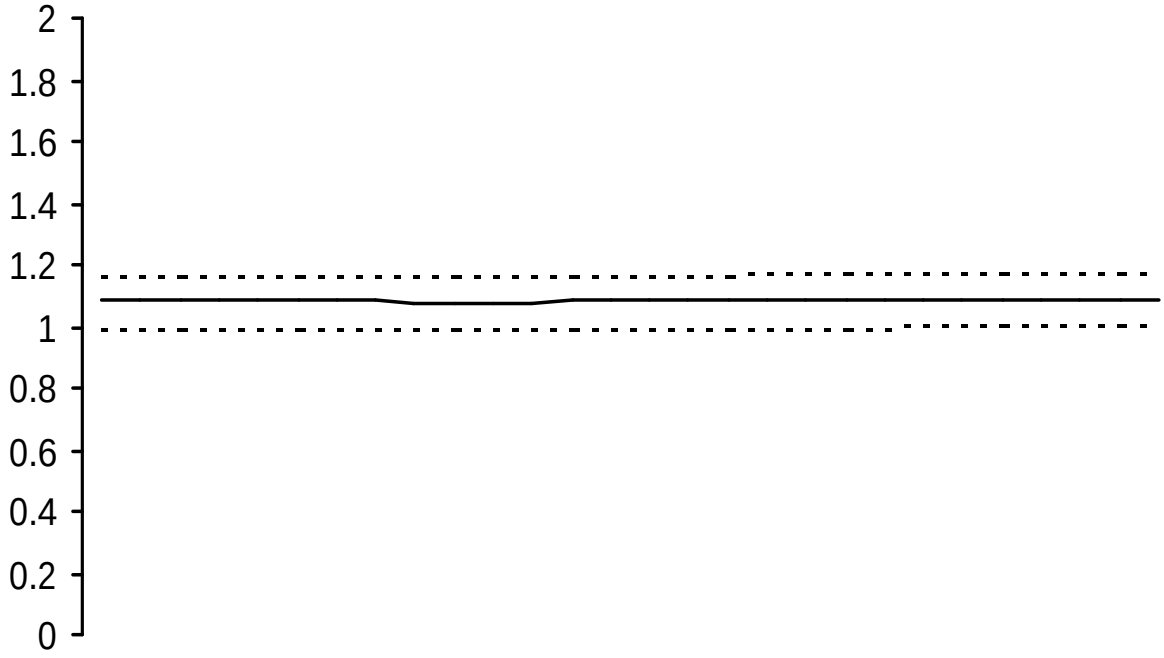


Figure 4.1: 5 yearly BT betas on daily data and the FTSE all share

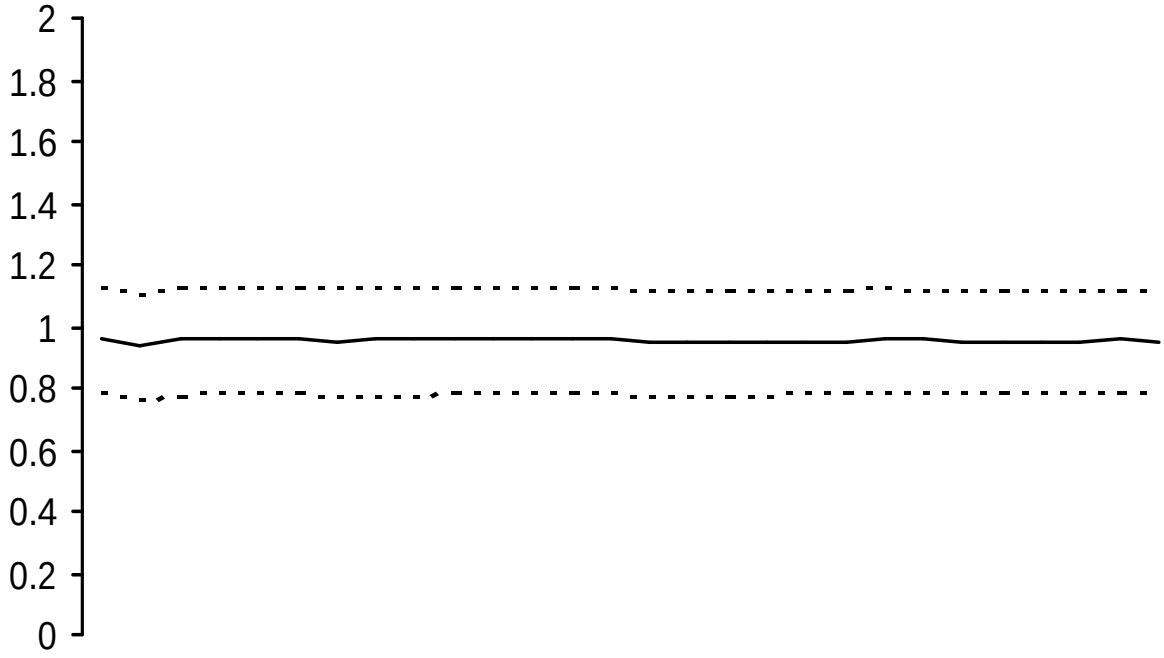


Figure 4.2: 5 yearly BT betas on weekly data and the FTSE all share

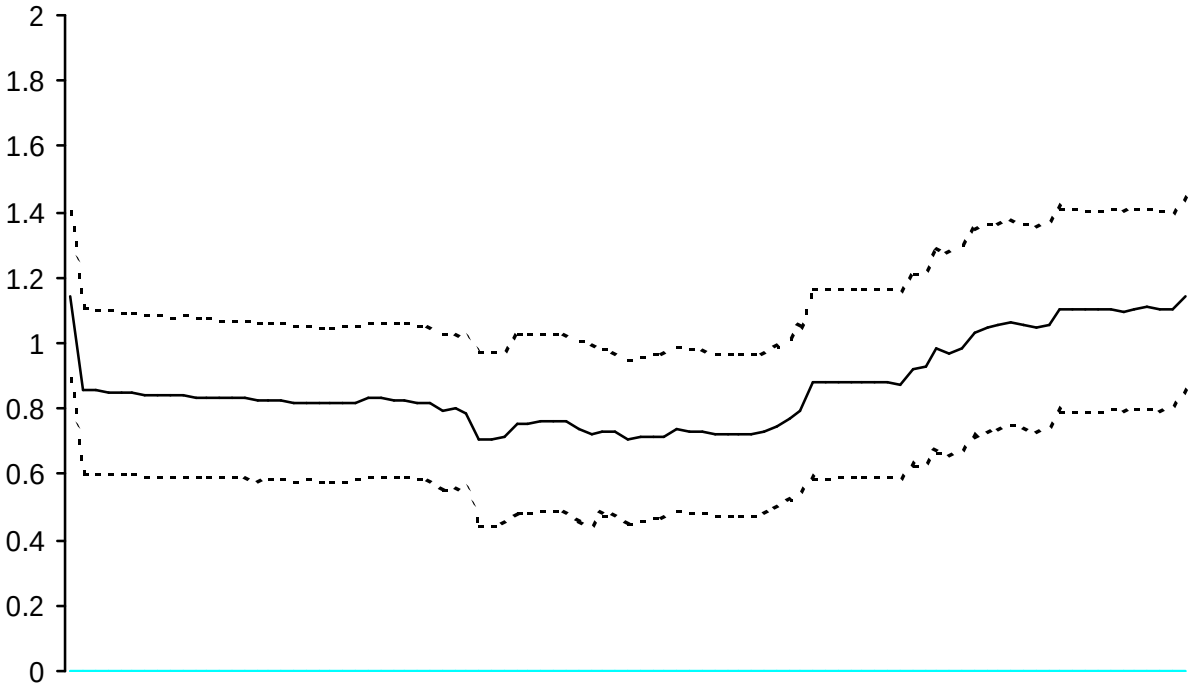


Figure 4.3: 5 yearly BT betas on monthly data and the FTSE all share

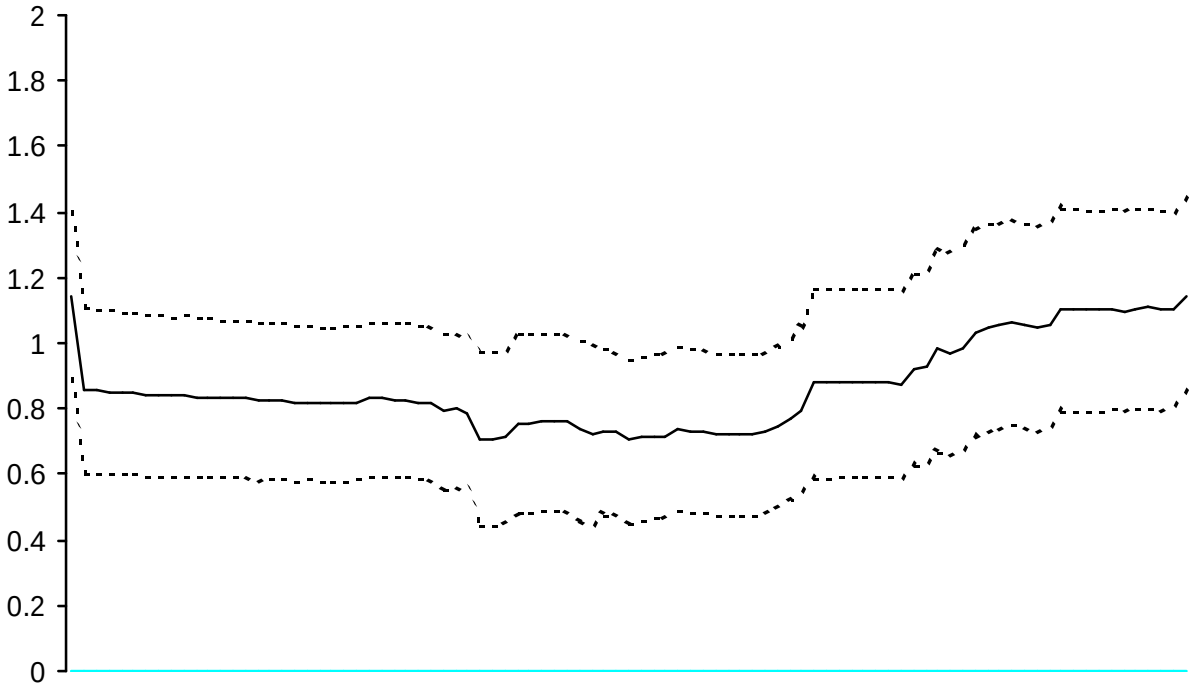


Figure 4.4: 5 yearly BT betas on quarterly data and the FTSE all share

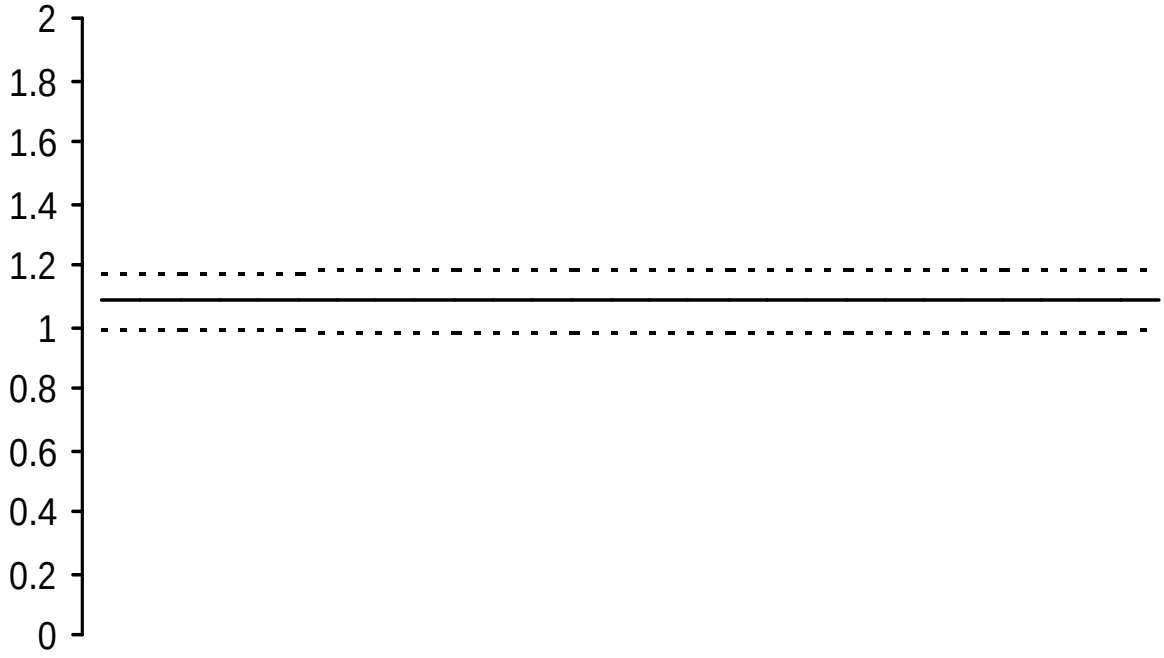


Figure 4.5: 5 yearly BT betas on daily data and a market index

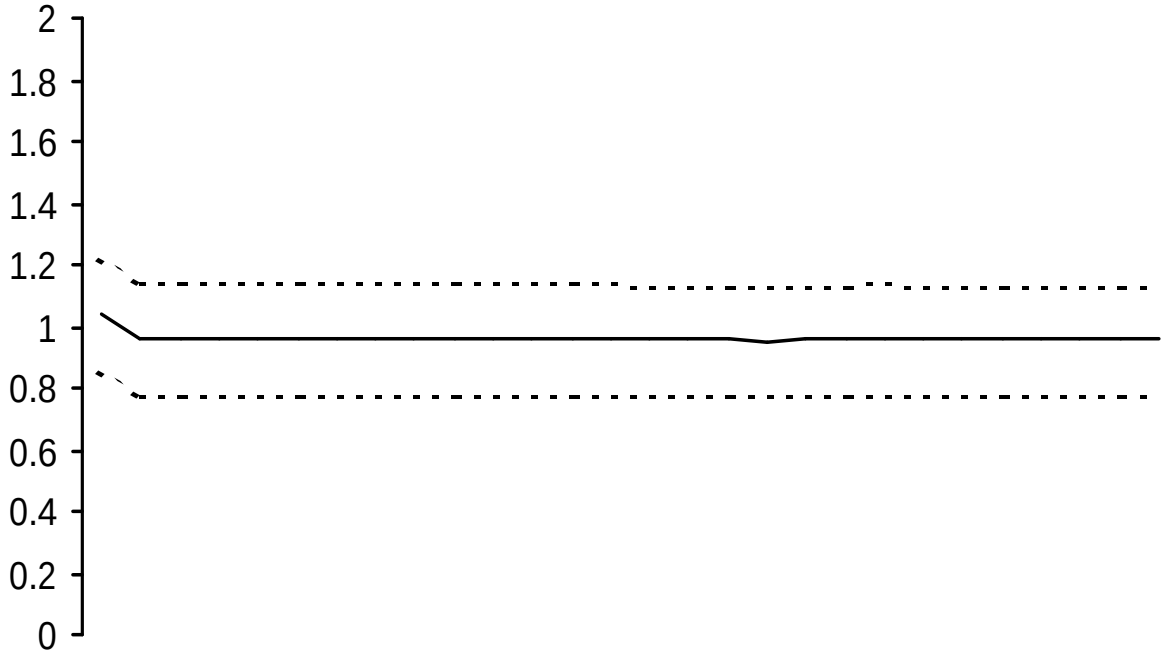


Figure 4.6: 5 yearly BT betas on weekly data and a market index

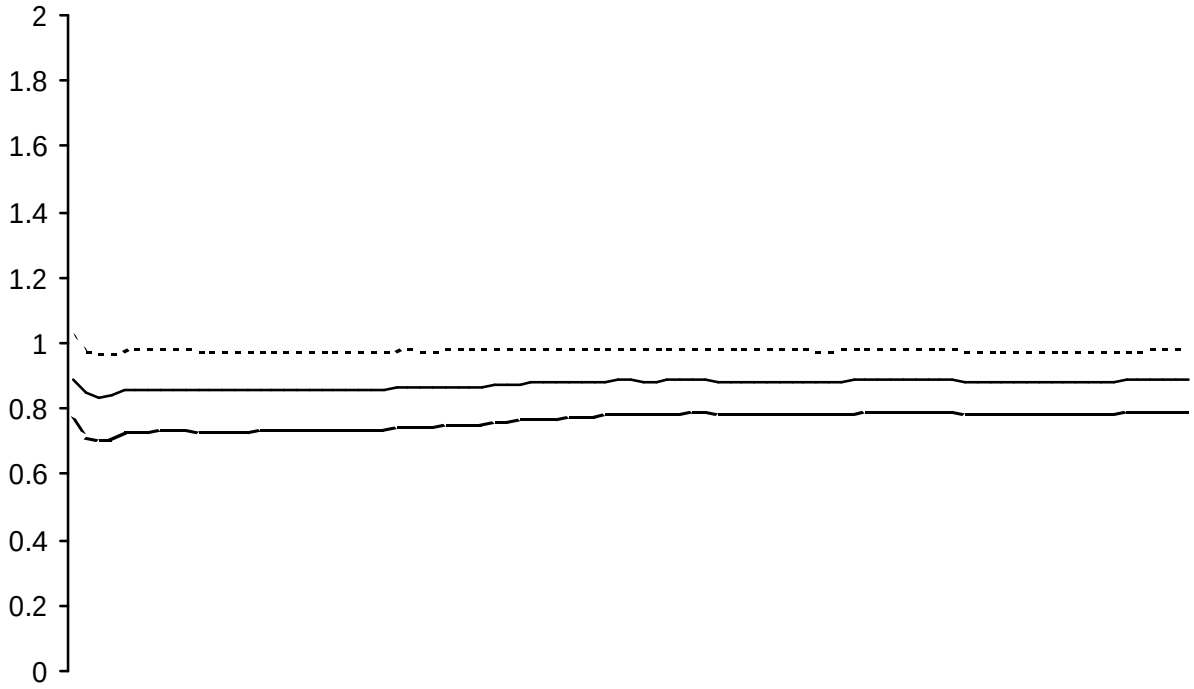


Figure 4.7: 5 yearly BT betas on monthly data and a market index

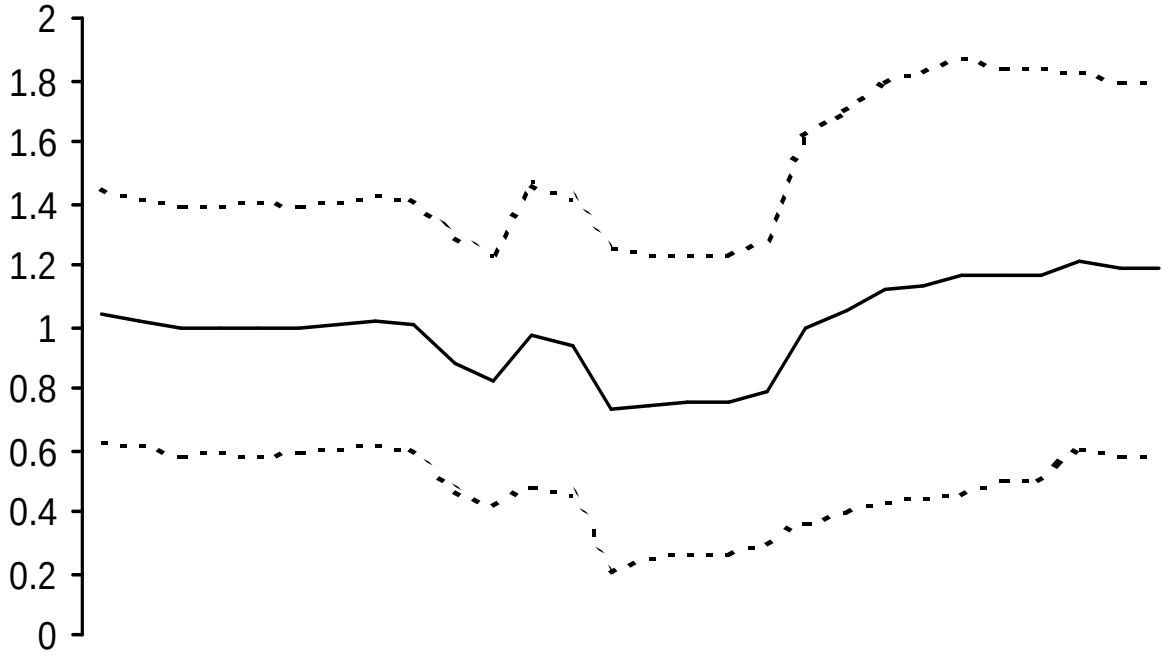


Figure 4.8: 5 yearly BT betas on quarterly data and a market index

For estimation purposes we should be measuring excess returns—that is returns in excess of the safe rate. Two issues arise. How exactly are returns measured and how do we measure the safe rate? The natural answer to the second question is that we should take the return on an asset that over the relevant period generates a return with (close to) a safe real rate of return. If the relevant time horizon for measuring returns is one day the asset is likely to be an overnight money market rate (eg LIBOR). With monthly data either monthly LIBOR or the return on one month Treasury bills could be used (and in practice it will not make much difference which is used). Although the real return on such assets is not certain, because inflation is not entirely predictable, with very short time horizons of under a month the divergence between the degree of certainty of real and nominal returns is small; with daily data it would only be an issue in times of hyperinflation.

The great advantage of defining returns in logarithmic form is that aggregation for returns over varying horizons is then exact (ie. the one month log return is the sum of the logs of the daily returns and so on). Thus the recommended definition of the excess return at time t is:

$$R_{it} = \ln\left(\frac{P_t + D_t}{P_{t-1}}\right) - \ln(1 + R_{ft})$$

where P_t is the price of the stock at date t, P_{t-1} is the price of the stock at $t-1$, D_t is per share dividend paid at time t (in practice this means the date at which the stock goes "ex" rather than the date at which payment of dividends is actually made), and R_{ft} is the one period safe rate of interest at time t.

4.4. Choice of market index, with particular focus on international mix of assets

What range of assets should be included in the market portfolio? The CAPM rests upon the mean variance approach. The key result there is that the market portfolio contains all the risky assets that exist *and* that all agents hold these risky assets in the same proportion

within their risky portfolios. It is obvious that these assumptions are strikingly at odds with the facts. The major holders of UK stocks are UK institutions (pension funds and life insurance companies). Their portfolios in recent years have been roughly 70% invested in assets issued by UK companies and by the UK government. Overseas assets make up only around 25% of all assets. Property, gold, paintings etc make up a smaller fraction of their portfolios than the value of such assets in global wealth.

One pragmatic approach is to take the CAPM as a guide and use as the market portfolio of risky assets a portfolio which reflects the composition of assets held by the dominant owners of the stocks in question. For most regulated UK companies this would imply the relevant portfolio is one with a high weight on UK equities (by which we mean FTSE all share stocks), a significant, but smaller, weight on UK bonds and with smaller weights on Continental European, Asian and US stocks. A portfolio that was 50% the FTSE all share index, 20% UK gilts, 10% overseas bonds and 20% the FTSE global share index (in £ terms) might be a rough approximation to this. (For more details on portfolios and the degree of international diversification see Dimson, Marsh, and Staunton (2001c); they estimate that UK investors put about 70% of their money into UK markets).

In practice when estimating betas it is more common to use the returns on an all equity portfolio and to use a domestic stock price index. The LBS risk measurement service uses the return on the FTSE all share index as its measure of the market return. Betas of many (though certainly not all) UK companies might be somewhat higher with respect to that portfolio than with respect to the "typical UK investor" portfolio outlined above. Some regulated companies with substantial overseas interests might have a higher beta on their overall activities with respect to the "typical UK investor" portfolio than with respect to a UK stock market index. The regulated UK activities of such international companies are likely, however, to have a higher beta with respect to a UK stock price index than with respect to a "typical UK investor" portfolio.

How to measure the beta of a part of the activities of a company—say the regulated UK activities of a company with a large overseas interest—is an important related issue. *If* we have a good estimate of the value of the relevant assets relative to the whole *and* we have estimates of the beta of the assets that are used in overseas business then a simple adjustment can be made to the overall company beta. Let the assets that generate regulated

business be denoted A_{reg} and the assets in other business be A_{oseas} Let the overall estimated beta for a company be β and assume we have a good estimate of the beta on overseas asstes which is β_{oseas}. The beta we want is the beta of the regulated activities, i.e., β_{reg}. A well know property of betas is that the beta on any portfilio is the weighted average of the betas on the constituent parts. This implies:

$$\beta = \beta_{reg} \times \frac{A_{reg}}{A_{reg} + A_{oseas}} + \beta_{oseas} \times \frac{A_{oseas}}{A_{reg} + A_{oseas}}$$

which can be re-arranged to yield:

$$\beta_{reg} = \beta \times \frac{A_{reg} + A_{oseas}}{A_{reg}} - \beta_{oseas} \times \frac{A_{oseas}}{A_{reg}}.$$

The practical problem, of course, is to estimate the relative value of the regulated to non-regulated parts of the business (i.e., $A_{reg}/(A_{reg} + A_{oseas})$) and the value of the beta on the non-regulated part, which we have called the "overseas" beta, β_{oseas}. A standard procedure is to estimate β_{oseas} by using stock price data on overseas companies in the same line of business and estimate their beta by reference to the market index used in constructing β. This market index we have called "the typical UK investor" and will have a high weight in UK equities. It is very likely that the resulting estimate of β_{oseas}will be well below the estimate of β so that the estimate of β_{reg} would be above β. If $A_{reg}/(A_{reg} + A_{oseas})$ were small the estimate of β_{reg} could be very much larger than β. For example, if regulated (UK) activities were 50% of all activities and the overall beta was 0.9 while the overseas beta was 0.3 the regulated Beta would be 1.5. Errors in measuring β_{oseas} can have a very major impact upon this calculation. Suppose that the true β_{oseas} was 0.5 rather than 0.3; the true value of β_{reg} would then be 1.3 rather than 1.5. If $A_{reg}/(A_{reg} + A_{oseas})$ were smaller than 1/2, the impact of mis-measurement of β_{oseas} would be amplified.

There is no easy answer to this problem. All one can do is to consider a range of different overseas companies as comparators and calculate a range of different estimates of β_{reg} based on the various alternatives and also based on different estimates of $A_{reg}/(A_{reg} + A_{oseas})$.

Of course the real issue for regulators is less about finding the beta of the assets associated with the UK activities of a company but more about finding an estimate of the beta of the

assets used in the regulated part of the business. If there are many assets used in the domestic business for activities which are not regulated then the overseas/domestic split is not the right one. But in practice the methodology for getting the appropriate beta is always going to be the same: we need an estimate of the overall company beta and then a separate estimate of the beta of the non-regulated bit. Then we use the well known relations above[5] to work out the beta of the regulated assets based on the relative weights of the assets used in the regulated part of the company to all its assets.

4.4.1. Empirical evidence

Table 4.3 shows estimates of the British Telecom beta based on 2 different estimates of the market return: the FTSE all share index (as in Table 4.1.2); a weighted average of 70% of the FTSE all share index and 30% of the FTSE world index (converted into sterling terms). Figures 4.1–4.8 show the rolling regression betas based on the different definitions of the market index.

The figures show that for daily or weekly data there is very little difference between the results. Things are rather different for the monthly data where using the mixed portfolio gives a much more stable estimate of beta which is lower than other estimates. Table 4.3 shows that the beta estimates based on the latest five years of data are little different with the "typical" (mixed) portfolio—this is largely because the typical portfolio has a large weight on UK stocks, itself a reflection of the so-called home bias puzzle.[6]

4.5. Bayesian Adjustments

The average beta across all stocks will be close to unity; if we use market weights in this averaging, and the market is the one used for estimating the betas, this average will be exactly unity. Betas on individual stocks will be estimated with error. It follows from these

[5] I.e., that the overall beta of a company is the weighted average of the beat of the assets used in different activities.

[6] The puzzle is why domestic investors hold such a high proportion of their wealth in domestic assets, thereby apparently missing out on opportunities for efficient portfolio diversification.

	No. of obs.	Beta: UK market index	Robust standard error	Beta: 70% UK; 30% world index	Robust standard error
Daily	1250	1.052	0.050	1.082	0.059
Weekly	260	0.960	0.088	1.066	0.223
Monthly	60	0.855	0.139	0.845	0.141
Quarterly	20	1.070	0.180	1.059	0.197

Table 4.3: Estimates of the beta of British Telecom—5 year regression window to August 2002

two propositions that for sampling reasons estimated betas significantly in excess of unity are likely to overstate beta and estimated betas well under unity are likely to be underestimates. A standard Bayesian adjustment to the "raw" estimate (eg the OLS estimate) takes account of the prior information (that the typical beta is unity). The resulting Bayesian estimate would be:

$$\beta_{adj} = \beta_{OLS} \times \frac{Var(\beta_{pop})}{Var(\beta_{pop}) + SE^2(\beta_{OLS})} + 1 \times \frac{SE^2(\beta_{OLS})}{Var(\beta_{pop}) + SE^2(\beta_{OLS})}$$

where: $SE^2(\beta_{OLS})$ is the standard error squared of the OLS estimate of beta (see equation (4.2) above), and $Var(\beta_{pop})$ is the variance of beta across the sample of firms for whom average beta is unity.

The logic behind the adjustment is straightforward. If there is lots of noise in estimating an individual stock beta from OLS regression ($SE^2(\beta_{OLS})$ is large) then one should attach a good deal of weight to the fact that on average we expect beta to be unity and relatively less weight to the estimate based on the OLS regression.

This equation assumes that the only information one has on the company beta besides the cross section average of unity is the OLS estimate—in other words prior to the OLS regression there was no reason to consider the most likely value of beta to be other than unity.

Reasonable estimates of $Var(\beta_{pop})$ can be assessed by looking at the dispersion of betas estimated by, for example, Bloomberg or by the LBS Risk Measurement Service. Both

Bloomberg and the LBS risk measurement service use their estimates of the variability across companies to make a Bayesian adjustment to their betas. Estimates of $SE^2(\beta_{OLS})$ are automatically produced by regression packages.

The logic behind making the Bayesian adjustment is strong. In practice, if daily data are used it is likely that $SE^2(\beta_{OLS})$ will be small relative to $Var(\beta_{pop})$ and the Bayesian adjustment will be small.

For example, the variance of the estimated betas of the FTSE 100 companies reported in the June 2002 edition of the LBS Risk Management service was around 0.13 [7]. (This is the variance of the estimated betas which have already been adjusted towards unity because LBS use a Bayesian adjustment). The estimates of the Vodafone beta in Table 4.2 are unadjusted. Table 4.4 below shows how those estimates would be adjusted using our estimate of the cross section variance of betas of 0.13.

	Unadjusted beta	Adjusted ("Bayesian") Beta
Five years monthly	0.99	0.99
Five years weekly	1.89	1.55
Five years daily	1.65	1.58

Table 4.4: Estimates of adjusted and unadjusted Vodafone Beta based on weekly, monthly and daily data

Unadjusted beats are as in Table 4.1.2 and are estimates of the Brattle Group (July 2002 report "Issues in Beta Estimation for UK Mobile Operators").

The important point about Table 4.4 is that even though the unadjusted estimate of the daily beta is very far from unity, the Bayesian adjustment has relatively little impact since the variance of the estimate of the Vodafone Beta based on daily data (0.13^2) is only one around one eighth the size of the variability in betas across companies.

[7] The highest estimated beta was 1.88 and the lowest 0.22.

4.6. Estimation of betas for companies with limited stock market data

There are several approaches to estimating the beta for the activities of a company where past data on the market rate of return on those assets derived from stock market prices is unavailable. The two common approaches are:

1. Use of the estimated betas of companies in similar lines of business and for whom stock market rates of return have been available.

2. Construction of an alternative beta estimate based upon the data that is available for the company in question. This data might include accounting rates of return or cash flow.

Strategy 1 is only successful where the comparator companies really are involved in the same line of business and have the same level of gearing. Adjusting for differences in gearing is straightforward and simply involves first un-gearing the beta of the comparator company to get an estimate of the asset beta and then using the current gearing of the company in question to re-gear the beta.[8] But there is no easy fix where the underlying business of the comparator companies is not really the same, either because they operate in different markets, produce different goods or are subject to different tax and regulations.

Strategy 2 depends upon being able to construct some sort of quasi rate of return from (usually) accounting data. One approach is to construct a rate of return by taking the ratio of accounting earnings to a measure of capital employed. A pure equity beta is, of course, the beta of the equity financed part of a firm's activity so an appropriate proxy based on accounting data could be:

$$\frac{\text{After tax earnings net of interest on debt (i.e., equity earnings)}}{\text{Total capital employed minus net debt}}. \tag{4.4}$$

Using this measure of the rate of return a beta can then be constructed in one of 2 ways.

[8] If the comparator company has a ratio of debt to debt plus equity (gearing) of g_1 and the gearing rate of the company for whom an estimate is required is g_0 the procedure is simply to multiply the comparator company equity Beta by the ratio $(1 - g_1)/(1 - g_0)$.

a by defining a time series of comparable accounting rates of return for all corporate assets using aggregate corporate sector counterparts for the numerator and denominator of equation (4.4). This series is then used as the proxy for R_{mt} in a regression equation analogous to equation (4.1) above;

b by using the accounting rate of return in place of the market rate of return for company i in a standard CAPM but using the stock market return on the market portfolio (i.e., like equation (4.1) using something like the FTSE all share index or the return on the "typical UK investor" portfolio).

Approach b is likely to give low and downward biased estimates of true beta since stock market returns are much more variable that accounting rates of return.

But there is evidence from the US that strategy a generates sensible looking numbers for beta which are quite highly correlated with the stock market beta. For example Beaver and Manegold (1975) find that there is very significant correlation between the accounting betas and the standard market betas for a large sample of firms where both could be estimated.

In practice estimation of the accounting betas following strategy a. above will give average betas that are likely to be close to unity—which is desirable—whereas strategy b. will likely generate low average betas. But following strategy a. requires that we estimate accounting rates of return for a very large sample of companies and not just for the company whose beta we are interested in. This is because we need to estimate a proxy for the market accounting rate of return. This means we need accounting data on a large sample of companies over a long period. If we only have annual rates of return we would need at least 20 years to get a beta estimate. This means that problems of time variation in underlying betas are likely to be very severe. Use of quarterly accounting data would help a good deal here. Nonetheless the practical difficult of using accounting data to estimate betas are formidable.

4.7. Conclusions

- There is a case to be made for using daily, or perhaps weekly, data rather than monthly data in estimating beta. For a share where trading is not significantly thinner or thicker than for the market as a whole, using daily data has real advantages.

- But where there may be a lag between the impact some events have on a particular share and the market in general going to lower frequency data can help. If one had to use the same estimation frequency for a very large number of different companies there is an argument that you have to go to weekly or monthly data because some stocks really take time to catch up with general market news.

- But regulators do not need to use the same frequency of data for estimates of different companies (unlike a commercial provider like the LBS which has standardised procedures and runs an automated service where all companies betas are calculated in the same way using 50 monthly observations). We conclude that using daily date may be right for many—but not—all companies.

- Adjusting standard errors for heteroskedasticity and serial correlation is important. Fortunately this is now a standard option in most econometric packages.

- A case can be made that a portfolio which reflects the mix of assets of the typical stock holder in the company should be used as the "market portfolio". For large UK companies whose shares are largely held by UK investors this implies a market portfolio with about 70% of its weight on UK assets and 30% on overseas assets. All returns should be in sterling.

- While in theory making a Bayesian adjustment is correct, in practice this may not make much difference if daily data is used because the standard error of the estimated beta is likely to be small relative to the variability in betas across companies. With monthly date the Bayesian adjustment is likely to be more significant.

5. Consistency in Cost of Capital Estimation

Even the best current estimates of the core components of asset pricing models are subject to considerable uncertainty. For example, standard errors on estimates of the equity risk premium are typically around 3%, on a point estimate of about 6.5%. This raises the key issue of how to deal with the inherent uncertainty surrounding cost of capital estimation.

One approach is simply to use the point estimate, accepting that approximately half the time the actual value of the cost of capital will be above, half of the time below this level. Such an approach implicitly assumes that the 'loss' from over-estimation is about the same as the 'loss' from under-estimation. Whether this assumption is valid depends very much on what the 'loss' is.

In cost of capital calculations for regulation, the loss from incorrect estimation takes the form of inefficient price setting and investment decisions. If the price cap is set too low (i.e., the cost of capital is underestimated), then the firm will make an inefficiently low investment in the market, which leads to a deadweight loss. If the price cap is set too high (i.e., the cost of capital is overestimated), then the regulator fails to restrain the exercise of market power by the monopolist; the deadweight loss in this case arises from price being set above marginal cost.

In order to show examine these factors in more detail, it is helpful to develop an explicit microeconomic model of price regulation, investment and cost of capital uncertainty. While the model is, inevitably, stylized, it captures the main issues that are of interest.

5.1. A Model of Price Cap Setting with Cost of Capital Uncertainty

Suppose that a regulator uses price cap regulation to regulate a monopolist: that is, the monopolist is constrained to charge a price no higher than $p*$, say. The monopolist sells a single product into a market with total demand given by $D(p)$, where p is the price set by the monopolist and $D(p)$ is the demand at that price. We assume (non-controversially) that demand is decreasing in price.

The monopolist has two costs to producing an amount q. The first is a production cost, relating to purchasing of equipment, wages paid to workers, the price of inputs etc.. Let this cost be $C(q)$. The second is the cost of capital: dividend payments on equity and interest payments on debt. Investors require a return of $k > 0$, say, on their investment, which is taken to be the production cost of the firm $C(q)$. Hence the total cost of the firm, which is the sum of production and financial costs, is $(1+k)C(q)$ when producing an amount q. In the rest of the analysis, we do not need to make any stronger assumptions on the cost function other than it is increasing and convex (and even the latter can be relaxed somewhat). It will make the exposition much clearer, however, if we specialize to the case in which marginal costs are constant and fixed costs are zero:

Assumption 3: $C(q) = cq$, *where* $c > 0$.

Here, c is the marginal cost of production. To repeat for emphasis: assumption 3 is made for convenience, and the following can be shown to hold for more general cost functions.

Finally, we assume that the regulator aims to maximize an unweighted sum of consumer and producer surplus; or, equivalently, to minimize the dead-weight loss in this market. The firm maximizes its profit.

5.2. The Deadweight Loss when the Cost of Capital is Known

There are two factors that a regulator must balance when setting the price cap. If the price cap is set too low, below the marginal cost of the regulated firm, then the firm will not

operate—if it did, it would make a loss. There is then a deadweight loss from the non-operation of the firm. It is worth clarifying at the outset what is meant by "not operating". One interpretation is that, literally, the regulated firm will exit its industry due to strict regulation. An alternative, less extreme interpretation is that this model analyses a specific investment or project, rather than an entry/exit decision; and that non-operation means that the project is not carried out. If the price cap is set too high, then the regulator fails to restrain the exercise of market power by the monopolist; the deadweight loss in this case arises from price being set above marginal cost.

The deadweight loss is plotted against the price cap in figure 5.1, assuming that the firm's cost of capital is known. Note that there are three regions. In region 0, the price cap is below the firm's marginal cost, and so the firm does not operate. The deadweight loss in this region is denoted DWL_0. If in this case the firm ceases operation altogether, then DWL_0 is the entire surplus that could have been gained from efficient operation of the firm i.e., the entire area under the demand curve:

$$DWL_0 = \int_{(1+k)c}^{p_0} D(p)dp$$

where p_0 is the 'choke-off' price at which demand is zero ($D(p_0) = 0$). In the case that the firm's response to the low price cap is to undertake an alternative, less valuable project, the deadweight loss will be the difference in the surpluses of the project that is cancelled and the substitute project.

In region I, the price cap is above marginal cost, but below the profit-maximizing price, p_m. In this region, the deadweight loss is increasing and convex in the price cap:

$$DWL_1 = \int_{(1+k)c}^{\overline{p}} D(p)dp - (\overline{p} - (1+k)c)D(\overline{p});$$
$$\frac{\partial DWL_1}{\partial \overline{p}} > 0; \qquad \frac{\partial^2 DWL_1}{\partial \overline{p}^2} > 0.$$

It is illustrated in figure 5.2. Finally, in region II, the price cap is above the profit-maximizing

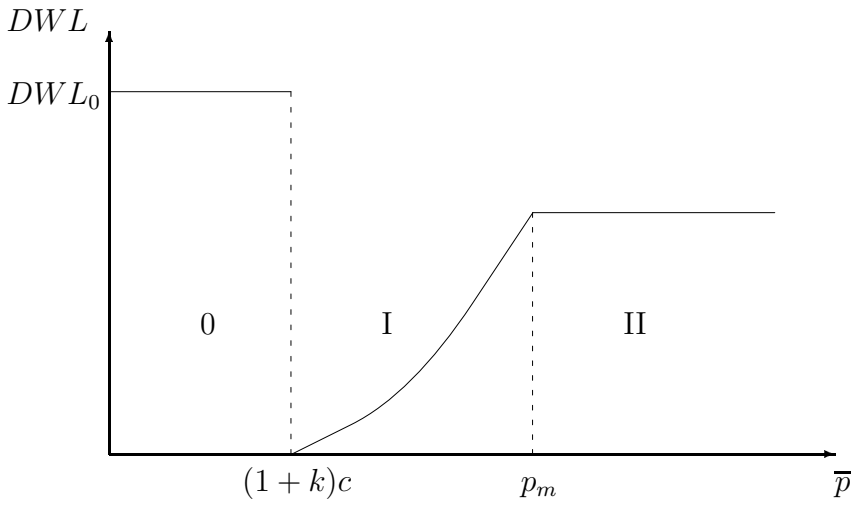


Figure 5.1: The Deadweight Loss against the Price Cap $\overline{p}$

price; the cap does not bind, and the deadweight loss is

$$DWL_2 = \int_{(1+k)c}^{p_m} D(p)dp - (p_m - (1+k)c)D(p_m).$$

This case is similar to region I, and so is not illustrated.

The optimal price cap when the firm's cost of capital is known by the regulator is straightforward—it equals the marginal cost $(1 + k)c$. In the next section, we consider how the optimal price cap changes when the regulator does not know the firm's cost of capital.

5.3. The Deadweight Loss when the Cost of Capital is not Known

Now suppose that the regulator is not fully informed about the firm's cost of capital, but must instead form an estimate of it that is subject to error. Suppose that the regulator's point estimate is $\hat{k}$; we model uncertainty in the following way:

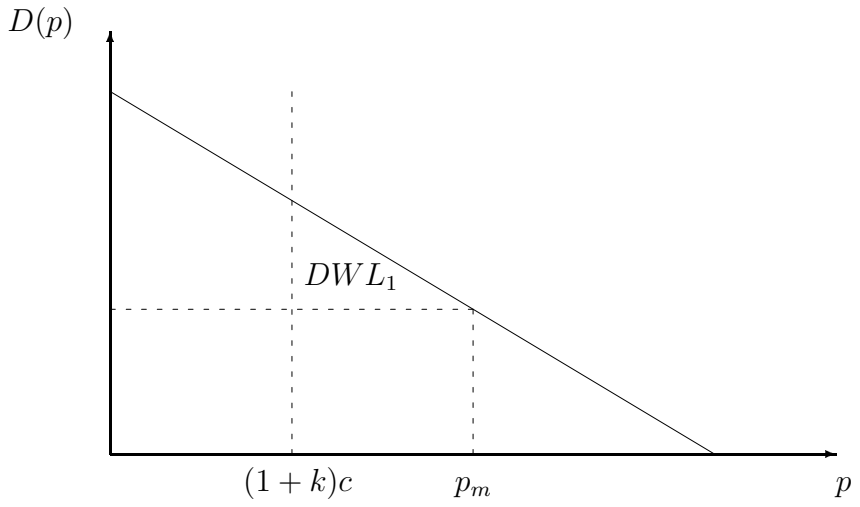


Figure 5.2: The Deadweight Loss in Region I

ASSUMPTION 4: *The regulator's estimate $\hat{k}$ equals the firm's true cost of capital k plus noise i.e., $\hat{k} = k + \eta$, where η is uniformly distributed on the interval $[-\varepsilon, \varepsilon]$ for $\varepsilon \geq 0$.*

The parameter ε therefore represents the accuracy of the regulator's estimate of the cost of capital. If $\varepsilon = 0$, then the regulator estimates perfectly the cost of capital; otherwise, there is error in the estimate. The 95% confidence interval attached to an estimate $\hat{k}$ is $[\hat{k} - 0.95\varepsilon, \hat{k} + 0.95\varepsilon]$. Both the regulator and the firm are perfectly informed about all other aspects: demand and production costs.[1] The assumption that the error is uniformly distributed is not at all restrictive: the uniform distribution simplifies calculations, but could easily be replaced with a more general distribution with little change to the qualitative conclusions.

[1]In this model, therefore, there is asymmetric information about the firm's marginal costs, as is the case in the seminal article on regulation, Baron and Myerson (1982). Note that the price-cap regulation considered here is not, in fact, the optimal regulatory form: in this set-up, that would typically require the regulator to give lump-sum transfers to firms. Such transfers are not usually observed, however, while price cap regulation (in one form or other) is commonly used e.g., in the U.K..

With this framework, we would like to address the question: should the regulator simply use the point estimate $\hat{k}$ of the cost of capital, or would it be better to use some other value? If some other value should be used, what factors determine the optimal value?

When the regulator estimates the cost of capital to be $\hat{k}$, it knows that the true cost of capital lies somewhere in the interval $[\hat{k}-\varepsilon, \hat{k}+\varepsilon]$; but as far as it is concerned, the true cost of capital is a random variable, with a uniform distribution over this interval. Three cases are important:

1. $DWL_0 > DWL_1((1+\hat{k}+2\epsilon)c)$ i.e., the deadweight loss from non-operation is substantially larger than the deadweight loss when the firm operates under the price cap. The expected deadweight loss in this case is illustrated in figure 5.3. The figure makes clear that the optimal price cap in this case is $\overline{p}^* = (1+\hat{k}+\epsilon)c$ i.e., the price cap is set so that even the highest cost firm operates. The reason is that the deadweight loss DWL_0 from non-operation is so large that it is optimal for the regulator to set the price cap to avoid any possibility of non-operation, while (subject to this) minimizing the mark-up of the monopolist.

2. $DWL_1((1+\hat{k}+\epsilon)c) > DWL_0$ i.e., the deadweight loss from non-operation is small compared to the deadweight loss from the monopoly mark-up. The expected deadweight loss in this case is illustrated in figure 5.4. The figure makes clear that the optimal price cap in this case is $\overline{p}^* < (1+\hat{k})c$ i.e., the price cap is set below the expected marginal cost of the firm. The reason is that the deadweight loss DWL_1 from the monopolist's mark-up is so large (compared to the deadweight loss from non-operation DWL_0) that it is optimal for the regulator to lower the price cap to minimize the mark-up.

3. $DWL_1((1+k+\epsilon)c) \leq DWL_0 < DWL_1((1+k+2\epsilon)c)$ i.e., the intermediate case. The expected deadweight loss in this case is illustrated in figure 5.5. The figure makes clear that the optimal price cap in this case is greater than the expected marginal cost $(1+\hat{k})c$, but smaller than $(1+\hat{k}+\epsilon)c$. Neither source of deadweight loss is dominant; the optimal price cap balances the two, leaving a positive probability that the firm does not operate (since $\overline{p}^* < (1+\hat{k}+\epsilon)c$) and a positive probability that there is a monopoly mark-up (since $\overline{p}^* > (1+\hat{k}-\epsilon)c$).

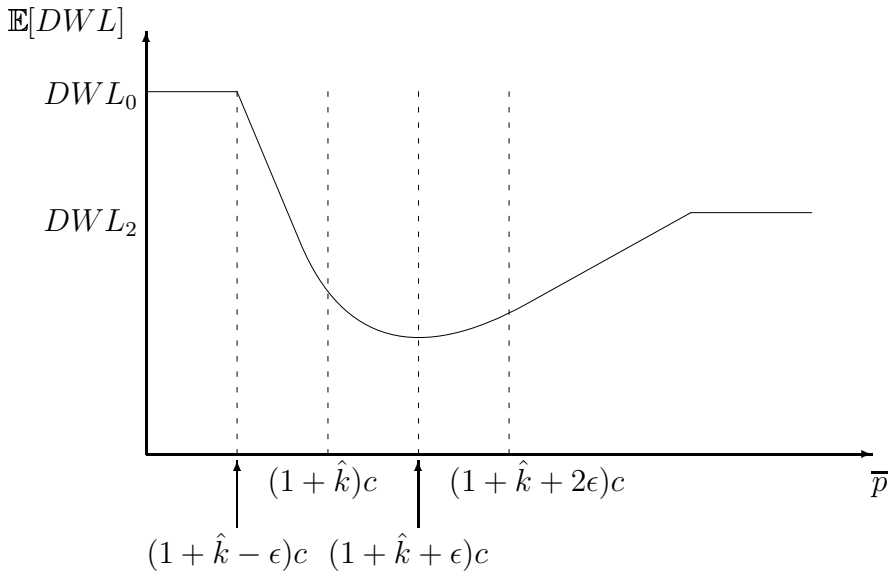


Figure 5.3: The Expected Deadweight Loss in Case 1

Note in particular that the price cap should be set equal to the expected marginal cost $(1+\hat{k})c$ (i.e., the marginal cost based on the point estimate of the cost of capital) only in the exceptional case that $DWL_0 = DWL_1((1+\hat{k}+\epsilon)c)$.

What can be said, then, about the optimal price cap $\overline{p}^*$? That depends on the case:

1. When the dead-weight loss from non-operation dominates, the optimal price cap is determined by the regulator's cost of capital estimate and the cost of the firm. The optimal price cap is higher when
   - the estimate of the cost of capital $\hat{k}$ is higher; and
   - the regulator's uncertainty about the cost of capital, ϵ, is higher.
2. When the dead-weight loss from non-operation is small, the optimal price cap is determined by the regulator's cost of capital estimate, the cost of the firm and market demand. The optimal price cap is higher when
   - the estimate of the cost of capital $\hat{k}$ is higher;

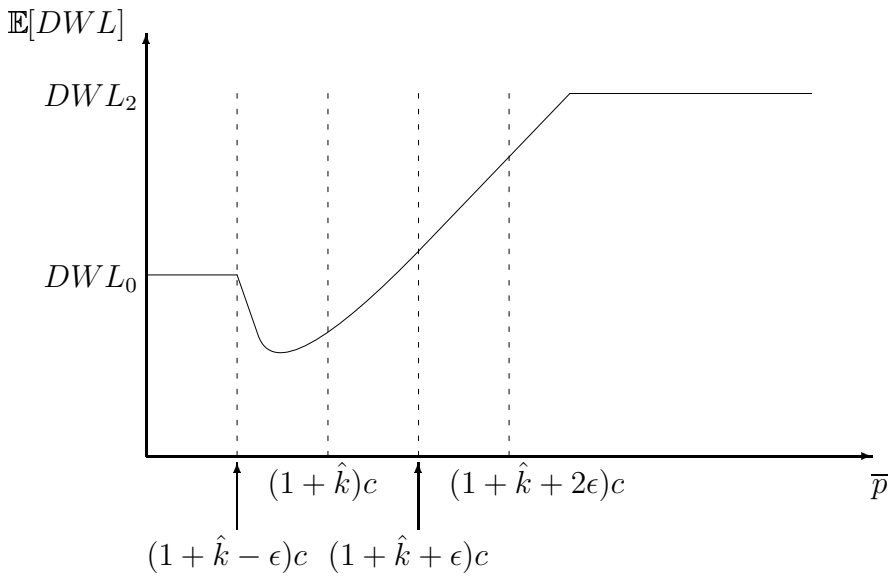


Figure 5.4: The Expected Deadweight Loss in Case 2

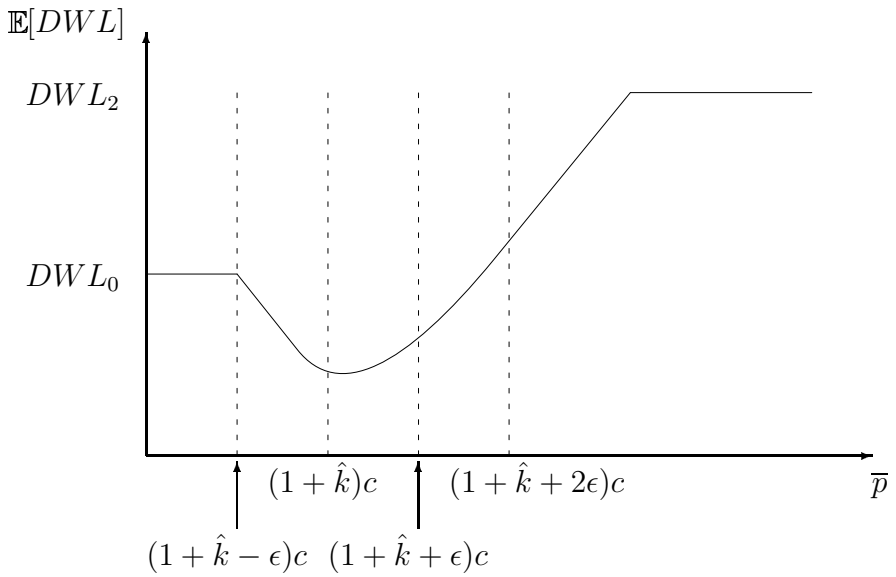


Figure 5.5: The Expected Deadweight Loss in Case 3

- the regulator's uncertainty about the cost of capital, ϵ, is lower;
- the deadweight loss from non-operation is higher;
- the elasticity of demand, measured by $D'(\overline{p}^*)$, is lower.

3. When the dead-weight loss from non-operation is moderate, the optimal price cap is again determined by the regulator's cost of capital estimate, the cost of the firm and market demand. The optimal price cap
   - is higher when the estimate of the cost of capital $\hat{k}$ is higher;
   - does not depend on the regulator's uncertainty about the cost of capital, ϵ;
   - is higher when the deadweight loss from non-operation is higher;
   - is higher when the elasticity of demand, measured by $D'(\overline{p}^*)$, is lower.

Some of the properties of the optimal price cap are obvious enough: for example, that it is higher when the point estimate of the cost of capital, or the deadweight loss from non-operation are higher. Some of the properties are more surprising. Consider the relationship between the degree of uncertainty and the optimal price cap. In case 1, greater uncertainty (a higher ϵ) *increases* the price cap, because in this case, the price cap is set equal to the estimate of the *highest* marginal cost. In case 2, however, greater uncertainty *lowers* the optimal price cap; this is because, in this case, the price cap is set with bearing in mind the estimate of the *lowest* marginal cost (in this case, the degree of the monopoly mark-up is all-important). In case 3, the degree of uncertainty has no effect on the optimal price cap: ϵ drops out of consideration when the two components of the expected deadweight loss are combined. Now consider the relationship between the optimal price cap and the elasticity of demand. When demand is elastic, the deadweight loss from the monopoly mark-up is large; hence the price cap should be low.

5.4. Extensions

Our analysis in this chapter is quite general—we have made few assumptions, and none of those has any significant qualitative effect on the conclusions. In this section, we consider briefly a couple of variations on the basic story.

5.4.1. Alternative Forms of Regulation

We have concentrated on cost of capital estimation under price cap regulation. This is a reasonable emphasis, given the prevalence of price cap regulation in the U.K.. Other forms of regulation are possible, however; and indeed other forms are used, even in the U.K.. The usual contrast to a price cap is rate of return regulation, where the level or rate of profits of the regulated firm is controlled. In its most extreme version, the firm would be regulated so as to achieve its target rate of return; it would, therefore, be allowed to increase prices in the event that costs are high. In this extreme case, there would be no dead-weight loss from non-operation of the firm; instead, inefficiency may arise from the low incentives that the regulated firm faces to decrease its costs.

While this is the usual contrast drawn between rate of return and price cap regulation, it is, in fact, fairly spurious, at least in the simplest setting. Both forms of regulation, if set to reflect *actual* costs, ensure that the firm always operates, and provide insufficient incentives for cost reduction. Both forms of regulation, if set to reflect *estimated* costs, can give rise to cases in which a high cost firm does not operate. The analysis in this section is not, therefore, specific to price cap regulation, but applies to any form of regulation that is set on the basis of estimated, rather than actual, costs.[2]

5.4.2. Alternative Forms of Uncertainty

The analysis has focussed on uncertainty about the firm's cost of capital, for obvious reasons. The analysis could equally well apply to uncertainty about different aspects of the regulated firm's operating environment. The analysis assumed that demand and production costs are known perfectly by the firm and the regulator; the only informational asymmetry concerned the cost of capital. It would be straightforward to extend the analysis to incorporate alternative aspects of information asymmetry and uncertainty. This would not change the qualitative conclusions in any significant way. The dead-weight loss trade-off faced by the regulator would remain; and hence different regulators are likely to use different val-

[2]If the regulated firm knows its true cost of capital, then an optimal regulation scheme can be used to induce the firm to reveal this information. See footnote 1, however.

ues for cost of capital estimates even when there are other dimensions to uncertainty and information asymmetry.

5.5. Conclusions

This chapter has developed a micro-economic analysis of regulation when the regulator is imperfectly informed about a firm's cost of capital. It has demonstrated the following key points:

- The optimal price cap is set taking into account the point estimate of the cost of capital.
- The optimal price cap is higher when the point estimate of the cost of capital is higher, when the deadweight loss from non-operation is higher, and when the elasticity of demand is lower.
- The optimal price cap may not depend on the degree of uncertainty in the estimate of the cost of capital.
- When the optimal price cap does depend on the degree of uncertainty, it may be either an increasing or decreasing function of the degree, depending on which component of the deadweight loss is dominant in the regulator's problem. If non-operation of the firm causes the greater welfare loss, then higher uncertainty increases the optimal price cap. Conversely, if a monopoly mark-up causes the greater welfare loss, then higher uncertainty decreases the optimal price cap.
- The results can be phrased in an alternative way by defining the *effective cost of capital estimate* to be the level of the cost of capital at which marginal cost equals the price cap.[3] A higher price cap corresponds, therefore, to a higher effective cost of capital; and vice versa.

[3] So, with a price cap of $\overline{p}$, the effective cost of capital $\overline{k}$ is given by $\overline{p} = (1 + \overline{k})c$ i.e., $\overline{k} = \overline{p}/c - 1$.

- The analysis then means that the effective cost of capital estimate that should be used by a regulator will depend on demand and cost conditions, as well as the point estimate and error in cost of capital estimation.
- Therefore two regulators who share the same point estimate and confidence interval for the costs of capital for their regulated firms will, in general, choose different effective costs of capital for price cap purposes, to reflect the demand and cost characteristics of the firm that they regulate.

6. Regulatory Risk

In this section, we consider the impact that regulation can have on a regulated firm's cost of capital. Three main issues are analysed:

1. What is the effect, if any, of regulatory inconsistency on a firm's cost of capital?
2. In what ways can different forms of regulation affect a firm's cost of capital?
3. How will a firm react to regulation that affects its cost of capital?

Incentive regulation (such as RPI - X) has three major objectives. The first is to ensure that the regulated firm does not charge excessive prices, and so reduce social welfare. The second is to provide the firm with an incentive to operate at minimum cost, to invest appropriately and to innovate. The third is to reveal to the regulator information about the firm that is relevant for the regulatory scheme.

These three objectives are generally in conflict. If a regulator were perfectly informed (about e.g., current and future costs), it could simply set the level of prices required to maximize social surplus. No such regulator exists, however, and any regulatory scheme must give up surplus to the regulated firm in order to elicit information and efficient investment. Optimal price regulation, for example, balances the three objectives by setting prices so that (i) by the end of the price control period, the regulated firm earns a level of return close to its cost of capital; (ii) the firm can earn excess profits during the review period if it lowers its costs; and (iii) the firm reports accurately its costs to the regulator.

Laffont and Tirole (1986) and Laffont and Tirole (1993) have shown that the best pricing scheme that can be used is a combination of a price cap (a ceiling on the prices that the firm can charge) and a cost-plus component (allowing certain cost changes to be passed

on in increased prices). The price cap provides an incentive for cost efficiency. The cost-plus component provides the firm with an incentive to report truthfully its costs. Both components are set as low as possible, subject to the provision of adequate incentives, to maximize social welfare.

These theoretical principles are applied widely in U.K. regulation. RPI - X is a hybrid of price cap and cost-plus regulation, emphasizing dynamic incentives towards cost reductions. Other forms of regulation are, of course, used. Revenue caps place a limit on the total income of a firm (such as used for Northern Ireland Electricity). Rate-of-return regulation allows the regulated firm to earn an agreed rate of return on its capital; this regulation is less commonly used in the U.K., but was favoured previously in the U.S..

A question that has received relatively little attention is whether, in setting regulation to provide the best incentives for efficiency and information, the regulator might contribute to the amount of risk that a firm faces. If regulation increases risk and hence the regulated firm's cost of capital, then this fact must be incorporated in the setting of e.g., the price cap. If it is not, then the price cap may be set too low and the investment decisions of the regulated firm distorted.

To assess the validity of this argument, we must first derive a satisfactory definition of 'regulatory risk': in what ways can regulation contribute to risk? Armed with a correct definition, we can then analyse the full effect of regulation on risk, recognising how the regulated firm can react.

6.1. The Definition of Regulatory Risk

The first step is to develop a satisfactory definition and description of regulatory risk. The most obvious definition states that regulatory risk arises whenever regulation affects the cost of capital of the regulated firm. This definition fails, however, to distinguish between two conceptually different forms of regulatory risk. The first arises from factors that are external to the firm and the regulator (such as macro-economic shocks), but have an impact on the regulatory scheme employed (e.g., the level of a price cap, in the case of RPI - X regulation). The second arises from factors that are under the regulator's control, and the choice of which

is regarded as uncertain by the regulated firm and investors. This section deals with the latter; the former is analysed in more detail in the next section.

A common concern among those involved in regulation is that the regulator can itself introduce risk, through unpredictable or unjustifiable regulatory intervention, so raising the regulated firm's cost of capital, and leading to inefficient investment. For example, Paul Plummer, then Chief Economist at the Office of the Rail Regulator, argued in 2000 that

> "A[n] ... issue concerns Railtrack's lack of a right of appeal to the Competition Commission in relation to the periodic review. This makes it even more important for the Regulator to adopt a consistent methodology and to explain the reasons for his decisions. Even so, it could result in additional *perceived* regulatory risks. Since this is unlikely to be in the long-term interests of funders, operators or users of the railway, the Regulator has said that he would in principle support the introduction of an appeal mechanism through the Transport Bill." (Emphasis in the original; see Plummer (2000))

Oftel has used the idea of regulatory risk to support its preferred method of regulation for indirect access for mobile networks: in 1999, it argued that

> " ... retail-minus ... avoids a major change in the regulatory framework which increases regulatory risk." (See Oftel (1999).)

Finally, Ofgem has stated as an advantage of price cap regulation that

> "[o]nce set, price caps create predictability, reduce regulatory risk during their period of operation, and create incentives for suppliers to increase efficiency. However, in dynamic retail markets Ofgem has tended to revise price controls every one or two years, which seriously limits their usefulness in driving efficiency and increases regulatory risk." (See Ofgem (2002).)

By definition, regulatory risk exists if and only if it affects the regulated firm's cost of capital. The central message of asset pricing theory is that only factors that co-vary with the

market portfolio (in the Capital Asset Pricing Model, or CAPM) or portfolios/factors (in an Arbitrage Pricing Theory, or APT, model) in equilibrium affect a firm's cost of capital. Hence 'regulatory risk' arises only when the regulator's actions co-vary with the market portfolio(s). Any regulatory action that has an effect that can be diversified does not contribute to risk.

This is simply a statement of the economic argument behind any asset pricing model, such as the CAPM. It is worth dwelling on the point, however, since there is a considerable amount of confusion on the point. For example, Ergas, Hornby, Little, and Small (2001), in a submission to the Australian Competition and Consumer Commission (ACCC), argue that

> "Firm specific risk which does not contribute to the risk of the market portfolio (i.e., which is not systematic) is not priced by the CAPM, even if it cannot be mitigated by diversification" (section 5.1).

This concern is valid only if the return from investing in the regulated firm has zero covariance with the returns of the market portfolio (i.e., non-systematic) and all other assets (i.e., non-diversifiable). To make the point clear, consider a portfolio comprised of N assets, each of which has a random return, denoted $\tilde{r}_i$ for the ith asset; let the portfolio weight of the ith asset be ω_i, where $\sum_{i=1}^{N} \omega_i = 1$. Then the variance of this portfolio is, from standard calculations,

$$V^P(N) = \sum_{i=1}^{N} \omega_i^2 V_i + \sum_{i=1}^{N} \sum_{j \neq i} \omega_i \omega_j C_{ij}$$

where V_i is the variance of the ith asset's returns and C_{ij} is the covariance of the ith and jth assets' returns. Note that there are N variance terms, but $N(N-1)$ covariance terms. It is this fact that leads to the conclusion that only covariance with the market portfolio matters to a well-diversified investor. If the assets and portfolio are symmetric, so that $V_i = V_j = V,\ i \neq j$, $C_{ij} = C_{kl} = C,\ i \neq j, k \neq l$, and $\omega_i = 1/N\ \forall i$, then the variance of the N-asset portfolio is

$$V^P(N) = \left(\frac{1}{N}\right) V + \left(\frac{N(N-1)}{N^2}\right) C;$$

in the limit, as the number of assets in the portfolio becomes very large, the portfolio variance approaches C more and more closely.

Consider adding an additional asset to the symmetric, N-asset portfolio just considered; suppose that this extra asset has zero covariance with all other assets, and has variance equal to V. Then the new portfolio's variance would be

$$V^P(N+1) = \left(\frac{1}{N+1}\right) V + \left(\frac{N(N-1)}{(N+1)^2}\right) C.$$

Again, in the limit as the number of assets in the portfolio becomes very large, the variance of this portfolio tends to C. In short: there is a negligible effect on a well-diversified investor of investing in an asset with risk that is non-systematic and non-diversifiable.

This narrows the focus of the search for the meaning of regulatory risk to actions that do not have diversifiable effects, but do have systematic effects. To be more explicit, regulatory risk (in the sense that is of interest in this section) arises only when the regulator takes actions that cause the returns of the regulated firm to co-vary with the returns on the market portfolio(s). One example of this is when a regulator decreases a price cap in response to a macro-economic shock that increases the profit of a firm. Is this example likely to occur in practice? One good reason why it might relates to learning by the regulator. Suppose that the regulator does not know the true marginal cost of a monopolist that it is regulating. Suppose also that observation of the firm's profit does not reveal perfectly to the regulator the marginal cost (because, say, there is also uncertainty over the level of demand). As the regulator observes the firm's profit over time, it learns about the true marginal cost. Finally, the regulator sets e.g., a price cap to limit the extent to which the firm can exercise its market power. If the regulator could observe the firm's marginal cost perfectly, it would set the price cap equal to marginal cost; or, if there is a fixed cost to cover, to average cost. In either case, the fully-informed price cap is low (high) for a low (high) cost firm.

Consider what happens when there is a positive macro-economic shock that increases both the return on the market portfolio and demand for the regulated firm's product. For a given price cap, the effect of the shock is to increase the firm's profit. As a result of observing a higher profit, the regulator revises its beliefs about the firm's marginal cost: under any reasonable updating scheme, the regulator will attach a higher probability to the

firm's marginal cost being low. In response to the positive shock, therefore, the regulator will lower the price cap. As a consequence, the firm's return may co-vary negatively with the market.

This source of systematic regulatory risk is particularly acute when the regulator has a large amount of discretion, in terms of both the frequency with which and the degree to which the regulator can adjust the price cap. If the regulator can make large adjustments very frequently to the price cap, then there is considerable systematic regulatory risk. If, on the other hand, the regulator is constrained to make small changes infrequently to the price cap, then there is little systematic regulatory risk from this source.

6.2. The Interaction of Systematic Risk and Regulation

Much attention has been paid to the effect of regulatory schemes on firms' incentives toward cost reduction and investment. Relatively little has been said about the impact of regulation on a firm's cost of capital. A major difference between the schemes outlined above is the degree of risk to which they expose the regulated firm. For the sake of illustration, suppose that a firm's costs are, in part at least, uncertain; and that this cost risk is not diversifiable. A firm faced with a price cap is exposed significantly to the risk: should its costs go up (for reasons that it cannot control), it is unable to raise its price in response. A same, but less acute situation obtains for a firm regulated by a revenue cap. It is able partially to alter its prices should costs change, but is constrained by its revenue cap. A firm allowed full cost pass-through, or facing rate-of-return regulation, faces no risk in this situation.

The different regulatory schemes therefore have markedly different implications for the firm's risk exposure. Two questions are key in the examination of this 'regulatory risk':

1. for any level of exogenous and non-diversifiable risk in demand and costs, what is the effect of various regulatory schemes on the degree of (non-diversifiable) risk of a regulated firm?

2. what is the effect of various regulatory schemes on the regulated firm's choice of investment risk?

The first question is discussed in section 6.2.1, the second in section 6.2.2. In both sections, we use a model in which a monopolist is faced with systematic risk about its marginal cost of production. The monopolist chooses its price after it knows its marginal cost; regulation is set before cost is realized. In section 6.2.1, the monopolist's project is given; we analyse how the use of different forms of regulation (a strict price cap, and a price cap with varying degrees of cost pass-through) affect the beta of a regulated firm. We argue that the likeliest outcome is that regulation increases the firm's beta. In section 6.2.2, the monopolist chooses between different projects that are distinguished by the amount of systematic risk that they entail. We show that price cap regulation leads the firm to opt a for project with lower risk than it would in the absence of regulation. If a limited degree of cost pass-through is permitted by the regulator, the same conclusion holds. Only if the firm is allowed to pass through a sufficiently large degree of its costs will it choose a project with the same degree of systematic risk as an unregulated firm. In section 6.2.1.2, we investigate if the conclusions are altered when the firm faces demand, rather than cost, uncertainty.

At the outset of this discussion, we should emphasize that we are concerned only with shocks to the firm's environment that are correlated with returns on the market portfolio—that is, we consider only the non-diversifiable risk that a firm faces. Prime examples for the sources of such risk are macro-economic shocks that may hit a firm. For example, cost risk may arise when the price of inputs, such as oil, vary with the macro-economic environment. Alternatively, it may be that labour costs are correlated, through wealth and general equilibrium effects, with the return on the market portfolio. Similarly, demand risk can arise through wealth effects that relate the market return to the demand for the product of a specific firm.

6.2.1. Regulation and Risk for a Given Project

6.2.1.1. Cost Uncertainty Consider first a monopolist facing a demand curve given by $D(p)$, describing the level of demand $D(\cdot)$ when it sets its price at p; the demand function is assumed to have all the usual, convenient properties. The firm's costs of production are $cD(\cdot)$, where $c \geq 0$ is a constant marginal cost; there are no fixed costs (this simplifying assumption will be discussed and relaxed later). The firm's cost of capital is denoted k; we

suppose that the firm's entire cost of production must be financed from equity, and so total costs are $(1+k)cD(\cdot)$.

Suppose first that the firm is regulated with a price cap that requires that the firm's price be no greater than $\overline{p}$. There are then three cases, distinguished by two marginal cost levels $\overline{c}$ and $\underline{c}$:

- $c \geq \overline{c}$: the firm chooses not to operate;
- $\overline{c} > c \geq \underline{c}$: the firm operates and the price cap binds;
- $c < \underline{c}$: the firm operates and the price cap does not bind.

$\overline{c}$ is given by the price/marginal cost equality $\overline{p} \equiv (1+k)\overline{c}$. $\underline{c}$ is given by profit-maximizing price/price cap equality $p^*(\underline{c}) \equiv \overline{p}$. Note that in the first case, it need not be that the firm ceases operation altogether: this could refer to an investment opportunity that the firm does not undertake.

The regulated firm's profit function $\pi^R(c)$ has, therefore, three components: for low marginal cost ($c < \underline{c}$), it is the same as the maximized profit function $\pi^*(c)$ of an unregulated firm; for intermediate cost ($\overline{c} > c \geq \underline{c}$), it a linear function; and for high cost ($c \geq \overline{c}$), it is zero:

$$\pi^R(c) = \begin{cases} 0 & c \geq \overline{c}, \\ (\overline{p} - (1+k)c)D(\overline{p}) & \overline{c} > c \geq \underline{c}, \\ \pi^*(c) & c < \underline{c}. \end{cases}$$

The unregulated firm's profit maximization problem is $\max_p (p-(1+k)c)D(p) \equiv \pi^*(c)$. It is a standard result that $\pi^*(c)$ is a convex function of c:

$$\frac{d\pi^*(c)}{dc} = -(1+k)D(p^*(c)) < 0; \qquad \frac{d^2\pi^*(c)}{dc^2} = -(1+k)D'(p^*(c))\frac{dp^*(c)}{dc} > 0.$$

(The first inequality follows from the envelope theorem; the second from the facts that demand is downward sloping and that the profit-maximizing price is increasing in costs, given the standard demand assumptions.) The functions $\pi^*(c)$ (the solid line) and the

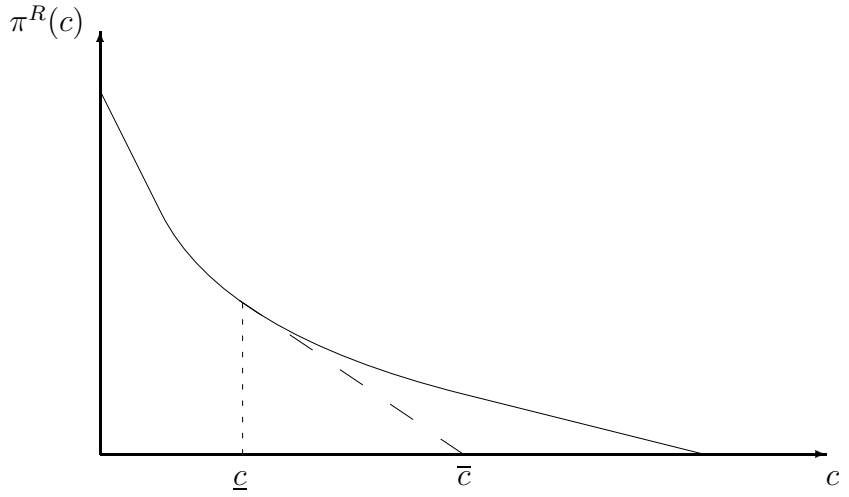


Figure 6.1: The profit functions $\pi^*(c)$ and $\pi^R(c)$

regulated firm's profit function $\pi^R(c)$ (the dotted line continuation) are shown in figure 6.1.

Now suppose that the marginal cost of the firm is uncertain: there is systematic risk from, for example, input prices that co-vary with the market. Hence c is a random variable, the value of which is realized before the firm chooses its price. The regulator, on the other hand, sets the price cap $\overline{p}$ before marginal cost is known. We suppose (for clarity) that the co-variance between marginal cost and the market portfolio return is -1, so that the beta of the firm is measured by minus the co-variance of profit with marginal cost:

$$\beta^* = -\text{Cov}\left[\pi^*(c), c\right], \qquad \beta^R = -\text{Cov}\left[\pi^R(c), c\right].$$

To allow explicit comparison of β^* and β^R, suppose that marginal cost can take one of two values, $c_L < c_H$, where the low cost realization occurs with probability $\phi \in [0, 1]$ and

the high cost realization with probability $1-\phi$. Straightforward calculations give

$$\begin{aligned}\beta^* &= 2\phi(1-\phi)\left(\pi^*(c_L)-\pi^*(c_H)\right)(c_H-c_L)\,,\\ \beta^R &= 2\phi(1-\phi)\left(\pi^R(c_L)-\pi^R(c_H)\right)(c_H-c_L)\,.\end{aligned}$$

It is unlikely that the price cap $\overline{p}$ will be set so that it never binds i.e., so that $c_H < \underline{c}$. It is also unlikely that the price cap will be set so that even the lowest cost firm does not operate i.e., so that $c_L > \overline{c}$. With these two restrictions, the shape of the two profit functions ensure that the beta of the regulated firm is greater than the beta of the unregulated firm i.e., $\beta^R > \beta^*$.

In practice, many firms (even those subject to price cap regulation) are allowed a degree of cost pass-through; for example, the RPI - X formula that is widely-used in the U.K. allows regulated prices to reflect general price inflation. This modification to price cap regulation is justified theoretically by the analysis of Laffont and Tirole (1986) and Laffont and Tirole (1993). Suppose, then, that the regulated firm's prices are constrained by a price cap of the form

$$\hat{p}(c) = \overline{p} + (1-\alpha)c$$

where $\alpha \in [0,1]$ is the extent of cost pass-through: when $\alpha = 0$, the regulated price $\hat{p}$ reflects fully realized cost, while when $\alpha = 1$, no cost pass-through is permitted.

There are two possible cases. Figure 6.2 illustrates the price cap when $\alpha > (1-k)/2$ i.e., cost pass-through is sufficiently incomplete. The unregulated, profit-maximizing price $p^*(c)$ and marginal cost $(1+k)c$ are also drawn. The figure makes clear that in this case, as before, there are two critical levels of marginal cost, $\underline{c}$ and $\overline{c}$, that define regions in which the price cap does not bind (when $c < \underline{c}$), in which it does bind and the regulated firm is willing to operate ($\underline{c} \leq c \leq \overline{c}$), and in which the regulated firm is not willing to operate ($c > \overline{c}$).

When cost pass-through is incomplete, therefore, the regulated firm's profit function is similar to the case analysed previously and illustrated in figure 6.1. The quantitative features of the profit function are changed by allowing cost pass-through; the qualitative features (the level and curvature of regulated profits relative to maximized profits) are unaltered.

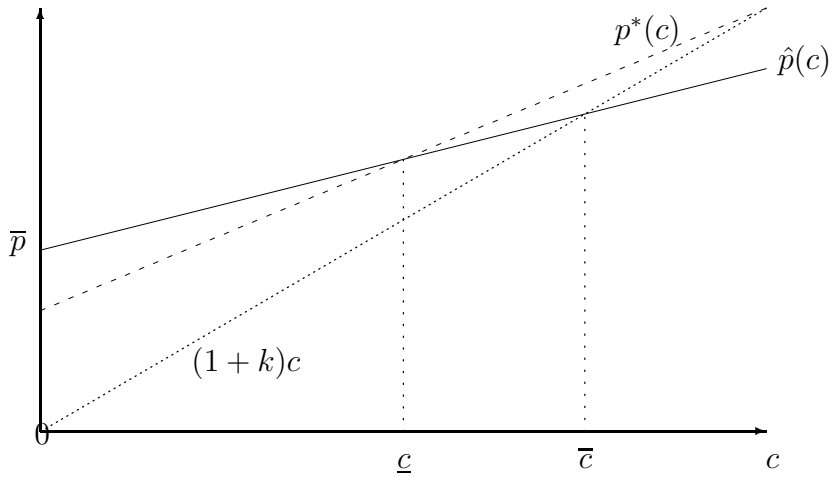


Figure 6.2: Price cap with incomplete cost pass-through

Consequently, the conclusion that the regulated firm's beta is greater than the unregulated firm's i.e., $\beta^R > \beta^*$ continues to hold even when partial cost pass-through is allowed. But the beta of the regulated firm is lower when partial cost pass-through is allowed.

If enough cost pass-through is allowed, however ($\alpha < (1-k)/2$), then the price cap does not bind, what ever the realized cost of the firm. In this case, the regulated firm's profit is clearly identical to the unregulated firm's; hence regulation does not affect the firm's beta. This is quite intuitive: with sufficient cost pass-through permitted by the regulator, the firm is never constrained by the price cap when choosing its price, whatever its marginal cost.

While the model that has been developed is specific in some of its details (for example, the cost structure of the firm), the conclusions are quite general. For example, fixed costs could quite easily be included without any major qualitative change in the conclusions.

6.2.1.2. Demand Uncertainty Consider a firm operating under a price cap (so that its price p must be less than or equal to some level $\overline{p}$) subject to demand shocks. Let the demand that the firm faces when it charges price p be $D(p;\theta)$, where $D(\cdot)$ is decreasing in

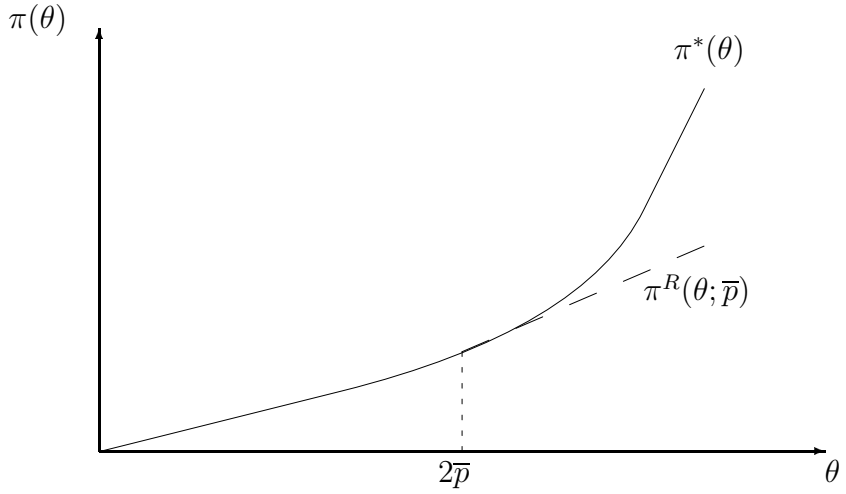


Figure 6.3: The profit functions $\pi^*(\theta)$ and $\pi^R(\theta;\overline{p})$

price, and θ is a random variable representing demand shocks. (This type of systematic risk can arise when the demand for the firm's product is driven by e.g., macro-economic shocks to income.) To make the point as simply as possible, suppose that $D(p;\theta) = \theta - p$. Finally, suppose that the firm has constant marginal cost of production and no fixed costs; without further loss of generality, let the marginal cost be zero.

An unregulated monopolist chooses price to maximize profit i.e., $\max_p(p-c)(\theta-p)$. Straightforward calculations show that the monopoly price is $p^*(\theta) = \theta/2$ and profit is $\pi^*(\theta) = \theta^2/4$. Now consider the regulated firm. For low enough values of θ, $\theta \leq 2\overline{p}$, the firm can choose its price to maximize its profit without being constrained by the price cap. But for larger values of $\theta > 2\overline{p}$, the price cap binds. Hence the profit of the regulated firm is

$$\pi^R(\theta;\overline{p}) = \begin{cases} \pi^*(\theta) & \theta \leq 2\overline{p}, \\ \overline{p}(\theta - \overline{p}) & \theta > 2\overline{p}. \end{cases} \tag{6.1}$$

Figure 6.3 plots the profit functions $\pi^*(\theta)$ (the solid line) and $\pi^R(\theta;\overline{p})$ (the dotted continuation) against the level of the demand shock parameter θ.

The figure demonstrates the same basic relationship between the regulated and unregulated profit functions as is observed when uncertainty originates from the cost side. In line with the previous argument, suppose that the co-variance between total demand and the market portfolio return is 1, so that the beta of the firm is measured by the co-variance of profit with the demand shock. The shapes of the two profit functions then ensure that the beta of the regulated firm is less than the beta of the unregulated firm. In the case of demand shocks, in contrast to cost uncertainty, the regulated firm's profit varies *less* than the profit of the unregulated firm. That translates into a *lower* beta.

6.2.1.3. Summary The simple model developed in this section has lead us to important conclusions:

- Price cap regulation affects the beta of the regulated firm.
- If uncertainty arises on the *cost* side, then price cap regulation *increases* the firm's beta. If uncertainty arises on the *demand* side, then price cap regulation *decreases* the firm's beta.
- Cost pass-through mitigates the effect of cost uncertainty. If sufficient cost pass-through is allowed, then the beta of a regulated firm is equal to the beta of an unregulated firm.

The models that have been developed in this section also allow us to make quantitative statements about situations under which price cap regulation will have a large effect on the beta of a firm. For example, consider the first case in which the firm's marginal cost is uncertain. Let

$$\Delta\beta \equiv \frac{\beta^R}{\beta^*} = \frac{\pi^R(c_L) - \pi^R(c_H)}{\pi^*(c_L) - \pi^*(c_H)}.$$

Hence $\Delta\beta$ is one measure of the change in a firm's beta effected by price cap regulation (normalizing by the unregulated firm's beta); it is greater than 1 in the cases of interest. In order to analyse this measure in greater detail, suppose that demand is linear: $D(p) = a - bp$,

where $a, b > 0$ are positive constants. Then

$$\pi^*(c) = \frac{(a - b(1+k)c)^2}{4b}.$$

When $c_L < \underline{c} < c_H < \overline{c}$,

$$\Delta\beta = \frac{(a - b(1+k)c_L)^2 - 4b\left(\overline{p} - (1+k)c_H\right)(a - b\overline{p})}{b^2(1+k)^2\left(c_H^2 - c_L^2\right)}.$$

Most obviously, $\Delta\beta$ is decreasing in $\overline{p}$: the more generous the price cap, the closer is the regulated firm to its unregulated counterpart. $\Delta\beta$ is increasing in a, the vertical intercept of the demand function. In words: the beta of a regulated firm operating in a large market (high a) will be greater than the beta of a firm in a small market. $\Delta\beta$ is increasing (decreasing) in b, the slope of the demand function, if $p^*(c_L) > (<)(1+k)c_H$ i.e., if the profit-maximizing price of the unregulated firm is greater (less) than the marginal cost of the high cost firm. Hence the beta of a regulated firm is non-monotonic in the value of the market: when the value is low, the beta is decreasing; when the value is high, the beta is increasing. Other comparative statics are similarly non-monotonic and determining the effects of an increase in a parameter requires that the values of parameters be specified.

More generally, armed with estimates of the demand and cost structures of the industry, and an estimate of the cost or demand uncertainty, this approach in principle allows the effect of regulation on a regulated firm's beta to be quantified.

6.2.2. Regulation and Project Choice

In this section, we consider what happens when the firm is able to choose its project. In order to shorten the exposition, we deal only with the case of marginal cost uncertainty; the case of demand uncertainty follows similar lines, as we will explain at the end of the discussion.

6.2.2.1. The Unregulated Firm As in section 6.2.1, we consider a monopolist facing a demand curve $D(p)$, with a constant marginal cost c, no fixed costs, and a cost of capital

of k. We now suppose that the firm has two choices to make. First, it decides which project to undertake. Projects are distinguished by their degree of systematic cost risk. Different projects have different probability distributions of marginal costs. We assume that all projects have the same expected marginal cost, but have different degrees of risk. Projects are, therefore, distinguished by mean-preserving spreads: let the distribution functions of marginal production costs for projects 1 and 2 be $F_1(c)$ and $F_2(c)$; then, for example,

$$\begin{aligned} \int_0^\infty c dF_1(c) &= \int_0^\infty c dF_2(c); \\ \int_0^y dF_1(c) &\geq \int_0^y dF_2(c), \quad y \geq 0. \end{aligned}$$

In fact, to simplify matters, we will assume that each project has two possible cost realizations, $c_L < c_H$, where the low cost realization occurs with probability $\phi \in [0,1]$ and the high cost realization with probability $1-\phi$. To ensure that each project has the same expected cost, we assume that $c_L = \phi$ and $c_H = 1+\phi$, so that the expected cost is $\hat{c} = \phi c_L + (1-\phi)c_H = 1$. The projects are distinguished, then, by their value of ϕ. Note that the projects with $\phi = 0$ or 1 are the least risky—they give a certain marginal cost of $\hat{c} = 1$. The beta of these projects is zero; the rate of return required from them is the risk-free rate R_f. Other projects are more risky i.e., have higher betas, but may have lower realized marginal cost; the riskiest has $\phi = 1/2$.

The second choice that the firm makes is to choose price to maximize profit. We assume that price is chosen after the project's marginal cost is realized. Hence the timing is

1. The firm chooses its project.
2. The project's marginal cost is realized.
3. The firm chooses its price.

The firm's maximized profit is therefore $\pi^*(c)$; see section 6.2.1 and figure 6.1.

The implication of convexity of the profit function is that, all other things equal, the firm will want to choose the riskiest possible project (i.e., the one with $\phi = 1/2$) in the prior stage. (This is a consequence of Jensen's inequality.) We suppose, however, that the

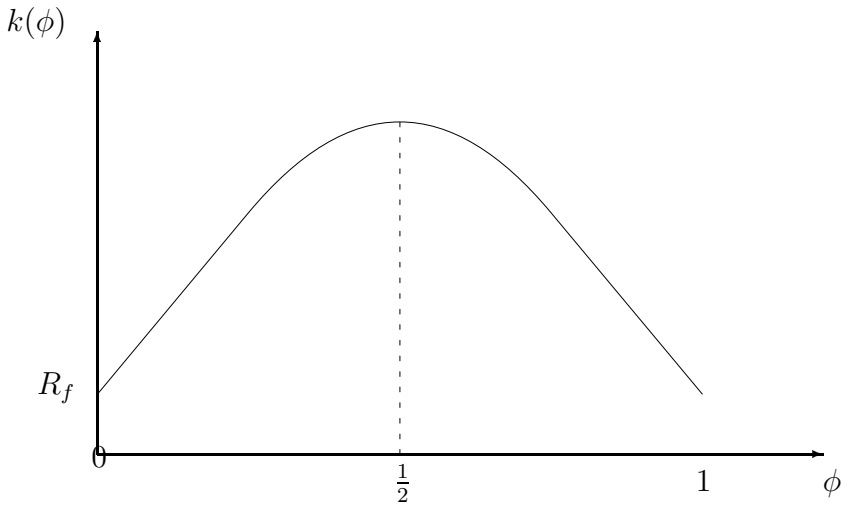


Figure 6.4: The cost of capital function $k(\phi)$

capital market can observe the firm's project choice; and that the firm's cost of capital k is a function of the project's risk, measured by ϕ. The cost of capital is a non-monotonic function of ϕ, reaching a maximum at $\phi = 1/2$ (the highest risk), and equalling the risk-free rate when $\phi \in \{0, 1\}$. This assumption is illustrated in figure 6.4.

The firm chooses the project i.e., ϕ to maximize expected profit:

$$\max_{\phi \in [0,1]} \mathbb{E}[\pi^*] \equiv \phi \pi^*(\phi; k(\phi)) + (1 - \phi)\pi^*(1 + \phi; k(\phi))$$

where we have emphasized that $k(\cdot)$ is a function of ϕ. The first-order condition for an interior optimum ϕ^* is

$$\frac{\partial \mathbb{E}[\pi^*]}{\partial \phi} = -\frac{\partial \mathbb{E}[\pi^*]}{\partial k} \frac{dk(\phi)}{d\phi}.$$

The left-hand side is positive because the profit function π^* is convex in cost. The right-hand side is positive in a region in which $dk/d\phi$ is positive (since expected profit is decreasing in the cost of capital). This equality shows how the firm balances at the margin the increase

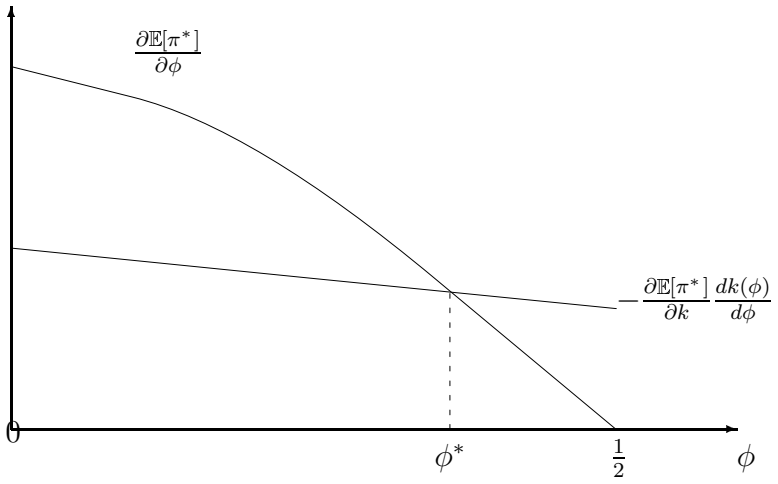


Figure 6.5: The unregulated firm's project choice ϕ^*

in expected profit from choosing a riskier project with the expected increase in the cost of capital from doing so. The trade-off is illustrated in figure 6.5.

6.2.2.2. The Firm Regulated by a Price Cap Now suppose that the firm is regulated with a price cap that requires that the firm's price be no greater than $\overline{p}$. See section 6.2.1 and figure 6.1 for the regulated firm's profit function $\pi^R(c)$. The firm's expected profit $\mathbb{E}[\pi^R]$ equals $\phi\pi^R(c_L) + (1-\phi)\pi^R(c_H)$. The regulated firm's profit-maximizing choice of project (assuming an interior solution) is given by

$$\frac{\partial \mathbb{E}[\pi^R]}{\partial \phi} = -\frac{\partial \mathbb{E}[\pi^R]}{\partial k}\frac{dk(\phi)}{d\phi}.$$

Since the price cap changes the curvature of the firm's profit function (see figure 6.1), it therefore changes the firm's choice of project, denoted ϕ^R. The price cap prevents the firm from raising its price to the profit-maximizing level if the realized cost is high. This means that the regulated firm gains less from a marginal increase in project risk (holding the cost

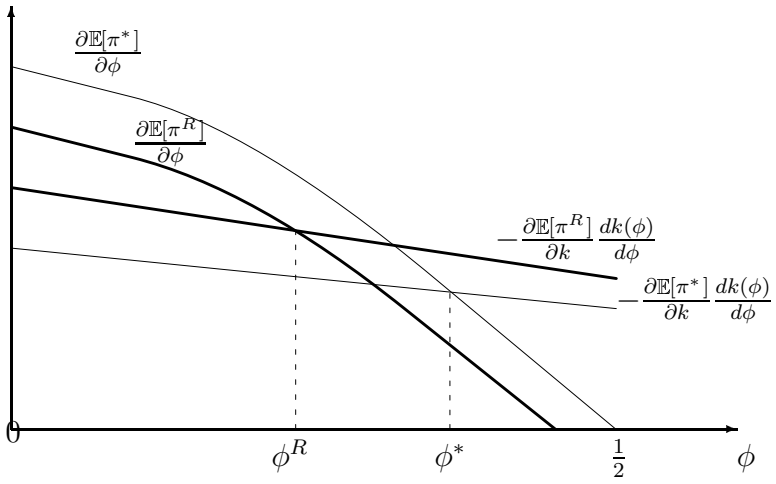


Figure 6.6: The regulated firm's project choice ϕ^R

of capital constant) than does the unregulated firm. It also means that the regulated firm's realized, and hence expected, profit falls by more than the unregulated firm's when the cost of capital increases. More explicitly,

$$\begin{aligned} \frac{\partial \mathbb{E}[\pi^*]}{\partial \phi} &\geq \frac{\partial \mathbb{E}[\pi^R]}{\partial \phi}, \\ -\frac{\partial \mathbb{E}[\pi^*]}{\partial k} &\leq -\frac{\partial \mathbb{E}[\pi^R]}{\partial k}. \end{aligned}$$

These two facts combined mean that $\phi^R < \phi^* < 1/2$: the regulated firm chooses a lower risk project than the unregulated firm. This is illustrated in figure 6.6.

6.2.2.3. The Firm Regulated with Cost Pass-through In section 6.2.2.2, we supposed that regulation takes the form of a strict price cap. Suppose instead that cost pass-through is allowed, so that the price cap takes the form

$$\hat{p}(c) = \overline{p} + (1 - \alpha)c$$

where $\alpha \in [0, 1]$ is the extent of cost pass-through.

The analysis in section 6.2.1 can be repeated, and the same two conclusions hold. When cost pass-through is incomplete ($\alpha > (1-k)/2$), the regulated firm's profit function is similar to the case analysed in section 6.2.2.2 and illustrated in figure 6.1. The quantitative features of the profit function are changed by allowing cost pass-through; the qualitative features (the level and curvature of regulated profits relative to maximized profits) are unaltered. Consequently, the previous conclusions continue to hold even when partial cost pass-through is allowed. If enough cost pass-through is allowed ($\alpha < (1-k)/2$), then the price cap does not bind, what ever the realized cost of the firm. Regulation does not affect the firm's choice of risk.

6.2.2.4. Summary In this section analysing the regulated firm's choice of project, we have dealt only with the case of cost uncertainty. In this case, we have found that a firm regulated with a price cap will choose lower beta projects; and that cost pass-through mitigates this effect. There is an equivalent conclusion for the case of demand uncertainty—a firm regulated with a price cap will choose higher beta projects.

6.3. Conclusions

We have demonstrated that price cap regulation changes the beta of a regulated firm when it is unable to choose its project. If uncertainty occurs on the cost side of a firm, then price cap regulation increases the firm's beta; if there is demand uncertainty, the regulation decreases the beta. Two things mitigate these effects. First, when the firm cannot choose its projects, an element of cost pass-through makes the effect of regulation less marked in the case of cost uncertainty. Secondly, when the firm is able to choose its project, and hence effectively the amount of uncertainty that it faces, its choice tends to reverse the effect of price cap regulation on its beta. These facts are all due to the change in shape of the firm's profit function that results from price cap regulation. While the arguments have been developed in simple settings (for example, the way in which different projects are modelled), the conclusions are, we believe, robust to a number of extensions.

The effects of regulation on a firm's beta therefore push in opposite directions, in the case of both cost and demand uncertainty. Which effect dominates? The answer to that question may depend on the time-scale. In the case of cost uncertainty, in the short-run, when the firm is unable to change its project, price cap regulation (provided cost pass-through is partial) increases the beta of the firm . In the medium- to long-run, the firm is able to choose its project; and so long-run betas should be lower then short-run betas. The converse holds for demand uncertainty—long-run betas should be higher than short-run betas.

REFERENCES

Andrews, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

Arbel, A., and P. J. Strebel (1983): "Pay Attention to Neglected Firms!," *Journal of Portfolio Management*, 9(2), 37–42.

Bansal, R., D. A. Hsieh, and S. Viswanathan (1993): "A New Approach to International Arbitrage Pricing," *Journal of Finance*, 48, 1719–1747.

Bansal, R., and S. Viswanathan (1993): "No Arbitrage and Arbitrage Pricing: A New Approach," *Journal of Finance*, 48(4), 1231–1262.

Banz, R. (1981): "The Relationship Between Return and Market Value of Common Stocks," *Journal of Financial Economics*, 9, 3–18.

Baron, D., and R. Myerson (1982): "Regulating a Monopolist with Unknown Cost," *Econometrica*, 50, 911–930.

Beaver, and Manegold (1975): "The Association Between Market Determined and Accounting Determined Measures of Systematic Risk: Some Further Evidence," *Journal of Financial and Quantitative Analysis*, 10(2), 231–284.

Black, F. (1972): "Capital Market Equilibrium with Restricted Borrowing," *Journal of Business*, 45, 444–454.

Black, F., M. C. Jensen, and M. Scholes (1972): "The Capital Asset Pricing Model: Some Empirical Tests," in *Studies in the Theory of Capital Markets*, ed. by M. C. Jensen, pp. 79–121. Praeger, New York.

Blanchard, O. J. (1993): "Movements in the Equity Premium," *Brookings Papers on Economic Activity*, 2(75–118).

Blume, M., and I. Friend (1973): "A New Look at the Capital Asset Pricing Model," *Journal of Finance*, 28, 19–33.

Breeden, D. (1979): "An Intertemporal Asset Pricing Model, with Stochastic Consumption and Investment Opportunities," *Journal of Financial Economics*, 7, 265–296.

Brennan, M., and Y. Xia (2002): "tay's as good as cay," Mimeo, Available at `http://finance.wharton.upenn.edu/~yxia/taynew.pdf`.

Brennan, M. J., and Y. Xia (2001): "Assessing Asset Pricing Anomalies," *Review of Financial Studies*, 14(4), 905–942.

Campbell, J. (1993): "Intertemporal Asset Pricing without Consumption Data," *American Economic Review*, 83, 487–512.

——— (2001a): "Understanding Risk and Return: the 2001 Marshall Lectures," Manuscript, Harvard University.

——— (2001b): "Why long horizons?A study of power against persistent alternatives," *Journal of Empirical Finance*, 8, 459–491.

Campbell, J., and R. Shiller (1988): "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," *Review of Financial Studies*, 1(195–227).

Campbell, J. Y., and J. H. Cochrane (1999): "By force of habit: A consumption-based explanation of aggregate stock market behavior," *Journal of Political Economy*, 107(205–251).

——— (2000): "Explaining the Poor Performance of Consumption-based Asset Pricing Models," *Journal of Finance*, 55, 2863–2878.

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay (1997): *The Econometrics of Financial Markets*. Princeton University Press, Princeton, New Jersey.

Chan, K. C., and N. Chen (1991): "Structural and Return Characteristics of Small and Large Firms," *Journal of Finance*, 46, 1467–1484.

Chen, N.-F. (1983): "Some Empirical Tests of the Theory of Arbitrage Pricing," *Journal of Finance*, 38(5), 1393–1414.

Chen, N.-F., R. W. Roll, and S. A. Ross (1986): "Economic Forces and the Stock Market," *Journal of Business*, 59(3), 383–403.

Clarida, R., J. Gali, and M. Gertler (1999): "The Science of Monetary Policy: A New Keynesian Perspective," *Journal of Economic Literature*.

Cochrane, J. (1997): "Where is the Market Going? Uncertain Facts and Novel Theories," *Economic Perspectives*, Nov/Dec.

Cochrane, J. H. (1996): "A Cross-sectional Test of an Investment-based Asset Pricing Model," *Journal of Political Economy*, 104, 572–621.

——— (2001): *Asset Pricing*. Princeton University Press, Princeton, New Jersey.

Connor, G., and R. A. Korajczyk (1993): "A Test for the Number of Factors in an Approximate Factor Model," *Journal of Finance*, 48(4), 1263–1291.

Constantinides, G., J. Donaldson, and R. Mehra (2002): "Junior can't borrow: a new perspective on the equity premium puzzle," *Quarterly Journal of Economics*.

Copeland, T., and J. F. Weston (1992): *Financial Theory and Corporate Policy*. Addison Wesley.

De Bondt, W. F. M., and R. M. Thaler (1985): "Does the Stock Market Overreact?," *Journal of Finance*, 40, 793–805.

Dhrymes, P., I. Friend, and M. N. Gultekin (1984): "A Critical Re-Examination of the Empirical Evidence on the Arbitrage Pricing Theory," *Journal of Finance*, 39(2), 323–346.

Dimson, E., P. Marsh, and M. Staunton (2001a): "Millenium Book II: 101 Years of Investment Returns," Discussion paper, London Business School.

——— (2001b): "Millenium Book II: 101 Years of Investment Returns," Discussion paper, London Business School.

——— (2001c): *Triumph of the Optimists*. Princeton University Press.

Ergas, H., J. Hornby, I. Little, and J. Small (2001): "Regulatory Risk," Mimeo, Available at `http://www.necg.com.au/pappub/papers-ergas-regrisk-mar01.pdf`.

Fama, E., and K. French (2001): "The Equity Premium," *Journal of Finance*.

Fama, E. F., and K. R. French (1988): "Dividend Yields and Expected Returns on Stocks and Bonds," *Journal of Financial Economics*, 22, 3–25.

——— (1992): "The Cross-section of Expected Stock Returns," *Journal of Finance*, 47, 427–465.

——— (1996): "Multifactor Explanations of Asset Pricing Anomalies," *Journal of Finance*, 51(1), 55–84.

Fama, E. F., and J. D. Macbeth (1973): "Risk, Return and Equilibrium: Empirical Tests," *Journal of Political Economy*, 81, 607–636.

Ferson, W. (1995): "Theory and Empirical Testing of Asset Pricing Models," in *Handbooks in Operations Research and Management Science*, ed. by R. Jarrow, V. Maksimovic, and W. Ziemba, pp. 145–200. Elsevier.

Ferson, W. E., and C. R. Harvey (1999): "Conditioning Variables and the Cross-section of Stock Returns," *Journal of Finance*, 54, 1325–1360.

Ferson, W. E., and R. W. Schadt (1996): "Measuring Fund Strategy and Performance in Changing Economic Conditions," *Journal of Finance*, 51, 425–461.

Ferson, W. E., and A. F. Siegel (2001): "Stochastic Discount Factor Bounds with Conditioning Information," Mimeo, Available at `http://www2.bc.edu/~fersonwa/AREA/opthj.pdf`.

French, K. (1980): "Stock Returns and the Weekend Effect," *Journal of Finance*, 8, 55–69.

Friend, I., and M. Blume (1975): "The Demand for Risky Assets," *American Economic Review*, 65, 900–922.

Friend, I., and R. Westerfield (1980): "Co-skewness and Capital Asset Pricing," *Journal of Finance*, 35, 897–913.

Gehr, A. (1975): "Some Testsof the Arbitrage Pricing Theory," *Journal of Midwest Finance Association*, pp. 91–105.

Gibbons, M., and P. Hess (1981): "Day of the Week Effect and Asset Returns," *Journal of Business*, 54, 579–596.

Giles, T., and D. Butterworth (2002): "Cost of Capital in the U.K. Mobile Services Market," Report to the Competition Commission on behalf of T-Mobile, Charles River Associates Limited.

Glassman, J., and K. Hassett (1999): *Dow 36,000: The New Strategy for Profiting from the Coming Rise in the Stock Market*. Times Books, New York.

Goetzmann, W., and P. Jorion (1999): "Global Stock Markets in the Twentieth Century," *Journal of Finance*.

Gordon, M. (1962): *The Investment, Financing, and Valuation of the Corporation*. Irwin, Homewood, IL.

Goyal, A., and I. Welch (2002): "Predicting the Equity Premium with Dividend Ratios," Working Paper 8788, NBER.

Graham, J. R., and C. R. Harvey (2001): "The Theory and Practice of Corporate Finance: Evidence from the Field," *Journal of Financial Econometrics*, 61, 1–28.

Greene, W. H. (1993): *Econometric Analysis*. Prentice Hall, Englewood Cliffs, second edn.

Hall, R. (2000): "'e-Capital': The Link Between the Stock Market and the Labor Market in the 1990s," *Brookings Papers on Economic Activity*.

Hall, S. G., D. K. Miles, and M. P. Taylor (1989): "Modelling Asset Prices With Time-Varying Betas," *The Manchester School*, LVII(4), 340–356.

Hansen, L. (1982): "Large Sample Properties of the Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Hansen, L. P., and R. Jagannathan (1991): "Implications of Security Market Data for Models of Dynamic Economics," *Journal of Political Economy*, 99, 225–262.

——— (1997): "Assessing Specification Errors in Stochastic Discount Factor Models," *Journal of Finance*, 52, 557–590.

Harrison, J. M., and D. M. Kreps (1979): "Martingales and Arbitrage in Multiperiod Securities Markets," *Journal of Economic Theory*, 20, 381–408.

Harvey, C. R., and A. Siddique (2000): "Conditional Skewness in Asset Pricing Tests," *Journal of Finance*, 55, 1263–1295.

Heaton, J. C., and D. J. Lucas (1999): "Stock Prices and Fundamentals," in *1999 Macroeconomics Annual*, ed. by B. Bernanke, and J. Rotemberg, Cambridge. MIT Press.

Huang, C., and R. H. Litzenberger (1998): *Foundations for Financial Economics*. North-Holland, New York.

Jagannathan, R., and Z. Wang (1996): "The Conditional CAPM and the Cross-section of Expected Returns," *Journal of Finance*, 51, 3–53.

——— (2001): "Empirical Evaluation of Asset Pricing Models: A Comparison of the SDF and Beta Methods," Working Paper 8098, NBER.

James, K. (2000): "The Price of Retail Investing in the UK," Occasional Paper Series 6, FSA.

Kan, R., and K. Q. Wang (2000): "Does the Nonlinear APT Outperform the Conditional CAPM?," Mimeo.

Kan, R., and G. Zhou (1999): "A Critique of the Stochastic Discount Factor Methodology," *Journal of Finance*, 54, 1221–1248.

——— (2001): "Empirical Asset Pricing: the Beta Method versus the Stochastic Discount Factor Method," Mimeo, Available at `http://www.rotman.utoronto.ca/~kan/research.htm`.

Keim, D. B. (1983): "Size Related Anomalies and Stock Return Seasonality: Further Empirical Evidence," *Journal of Financial Economics*, 12, 13–32.

Kiley, M. (2000): "Stock Prices and Fundamentals in a Production Economy," Finance and Economics Discussion Paper 05, Federal Reserve Board.

Kocherlakota, N. (1996): "The Equity Premium: It's Still a Puzzle," *Journal of Economic Literature*, 34, 42–71.

Kothari, S. P., J. Shanken, and R. G. Sloan (1995): "Another Look at the Cross-section of Expected Stock Returns," *Journal of Finance*, 50, 185–224.

Kraus, A., and R. H. Litzenberger (1976): "Skewness Preference and the Valuation of Risk Assets," *Journal of Finance*, 31, 1085–1100.

——— (1983): "On the Distributional Conditions for a Consumption-oriented Three Moment CAPM," *Journal of Finance*, 38, 1381–1391.

Laffont, J.-J., and J. Tirole (1986): "Using Cost Observation to Regulate a Firm," *Journal of Political Economy*, 94(3), 614–641.

——— (1993): *A Theory of Incentives in Procurement and Regulation*. MIT Press, Cambridge.

Lehmann, B. (1992): "Empirical Testing of Asset Pricing Models," in *New Palgrave Dictionary of Money and Finance*, ed. by P. Newman, M. Milgate, and J. Eatwell, pp. 749–759, New York. Stockton Press.

Lehmann, B. N., and D. M. Modest (1987): "Mutual Fund Performance Evaluation: A Comparison of Benchmarks and Benchmark Comparisons," *Journal of Finance*, 42(2), 233–265.

Leland, H. (1999): "Beyond Mean-Variance: Risk and Performance Measurement in a Nonsymmetrical World," *Financial Analysts Journal*, 1, 27–36.

Lettau, M., and S. Ludvigson (2001a): "Consumption, Aggregate Wealth and Expected Stock Returns," *Journal of Finance*, 56, 815–49.

——— (2001b): "Resurrecting the (C)CAPM: A Cross-Sectional Test when Risk Premia are Time-Varying," *Journal of Political Economy*, 109(6), 1238–1287.

——— (2002): "tay's as good as cay: Reply," Mimeo, Available at `http://www.econ.nyu.edu/user/ludvigsons/reply.pdf`.

Lim, K.-G. (1989): "Selection of Risky Investments in Stock Portfolios and Capital Budgets," *Journal of Financial and Quantitative Analysis*, 24, 205–216.

Lintner, J. (1965): "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *Review of Economics and Statistics*, 47, 226–37.

MacKinlay, A. C. (1995): "Multifactor Models do not Explain Deviations from the CAPM," *Journal of Financial Economics*, 38, 3–28.

Markovitz, H. (1952): "Portfolio Selection," *Journal of Finance*, 7, 77–91.

McGrattan, E., and E. Prescott (2001): "Is the Stock Market Overvalued?," Working Paper 8077, NBER.

Merton, R. C. (1973): "An Intertemporal Capital Asset Pricing Model," *Econometrica*, 41, 867–887.

Newey, and West (1987): "A Simple Positive Semi-definite Heteroskedasticity and Autocorrelation Consistent Covariance Estimator," *Econometrica*, 55, 703–708.

Ofgem (2002): "Review of domestic gas and electricity competition and supply price regulation: Conclusions and final proposals," Available at http://www.ofgem.gov.uk/docs2002/price_regulation_review.pdf.

Oftel (1999): "Oftel's review of the mobile market: Statement issued by the Director General of Telecommunications," Available at http://www.oftel.gov.uk/publications/1999/competition/mmrv799.htm.

Pickford, S., and Wright (2000): "The Equity Risk Premium, or Believing Six Nearly Impossible Things Before Breakfast," Report 145, Smithers & Co Ltd.

Plummer, S. (2000): "What Regulatory Risk?," Available at http://www.rail-reg.gov.uk/speeches/.

Reinganum, M. R. (1983): "The Anomalous Stock Market Behavior of Small Firms in January: Empirical Tests for Tax-Loss Selling Effects," *Journal of Financial Economics*, 12, 89–104.

Robertson, D., and S. Wright (2002): "The Good News and the Bad News about Long-Run Stock Returns," Mimeo.

Roll, R. (1977): "A Critique of the Asset Pricing Theory's Tests," *Journal of Financial Economics*, 4, 129–176.

——— (1983): "On Computing Mean Returns and the Small Firm Premium," *Journal of Financial Economics*, 12, 371–386.

Roll, R. W., and S. A. Ross (1980): "An Empirical Investigation of the Arbitrage Pricing Theory," *Journal of Finance*, 35(5), 1073–1103.

Ross, S. (1976): "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, 13, 341–360.

——— (1978): "Mutual Fund Separation in Financial Theory: The Separation Distributions," *Journal of Economic Theory*, 17, 254–286.

Satchell, S., and S. Hwang (1999): "Modelling Emerging Market Risk Premia Using Higher Moments," *International Journal of Finance and Economics*, 4(4), 271–296.

Scott, R. C., and P. A. Horvath (1980): "On the Direction of Preference for Moments of Higher Order than the Variance," *Journal of Finance*, 35, 915–919.

Sears, R. S., and K. C. J. Wei (1985): "Asset Pricing, Higher Moments, and the Market Risk Premium: a Note," *Journal of Finance*, 40, 1251–1253.

Shanken, J. (1990): "Intertemporal Asset Pricing: An Empirical Investigation," *Journal of Econometrics*, 45, 99–120.

Sharpe, W. F. (1964): "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *Journal of Finance*, 19, 425–42.

Shiller, R. (2000): *Irrational Exuberance*. Princeton University Press.

Siegel, J. J. (1998): *Stocks for the Long Run*. McGraw-Hill, second edn.

Smithers, A., and S. Wright (2000): *Valuing Wall Street*. McGraw-Hill.

——— (2002): "Stock Markets and Central Bankers: The Economic Consequences of Alan Greenspan," *World Economics*, 3(1).

Taylor, J. B. (1993): "Discretion vs Policy Rules in Practice," *Carnegie Rochester Series on Public Policy*, 39(0), 195–214.

Wadhwani, S. (1999): "The US Stock Market and the Global Economic Crisis," Discussion paper, National Institute of Economic and Social Research Review.

White (1980): "A Heteroskedasticity Consistent Co-variance Matrix Estimator as a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–838.