A Skeptical Appraisal of Asset-Pricing Tests

Jonathan Lewellen
Dartmouth and NBER
jon.lewellen@dartmouth.edu

Stefan Nagel
Stanford and NBER
nagel_stefan@gsb.stanford.edu

Jay Shanken
Emory and NBER
jay_shanken@bus.emory.edu

This version: January 2006
First draft: May 2005

We are grateful to Christopher Polk and workshop participants at Stanford and the universities of Alberta, Calgary, and Washington for helpful comments. We thank Ken French, Sydney Ludvigson, Stijn Van Nieuwerburgh, and Motohiro Yogo for providing data via their websites.

A Skeptical Appraisal of Asset-Pricing Tests

Abstract

The finance literature has proposed a wide variety of new asset-pricing models in recent years, motivated by evidence that small, high-B/M stocks have positive CAPM-adjusted returns. Indeed, it is now standard practice to evaluate a model based on how well it explains average returns on the Fama-French 25 size-B/M portfolios, something many models seem to do remarkably well. In this paper, we provide a skeptical appraisal of the empirical methods used in the literature. We argue that asset-pricing tests are often highly misleading, in the sense that apparently strong explanatory power (high cross-sectional R^2s and small pricing errors) in fact provides exceptionally weak support for a model. We offer a number of suggestions for improving empirical tests and evidence that several proposed models don't work as well as originally advertised.

1. Introduction

The finance literature has offered an explosion of new asset-pricing models in recent years, motivated by evidence that small, high-B/M stocks have positive CAPM-adjusted returns. The new models – formal equilibrium theories as well as simple econometric models – propose a variety of risk factors beyond the market return or aggregate consumption. Proposed factors include labor income (Jagannathan and Wang, 1996; Heaton and Lucas, 2000), growth in real investment, GDP, and future consumption (Cochrane, 1996; Vassalou, 2003; Li, Vassalou, and Xing, 2005; Parker and Julliard, 2005; Hansen, Heaton, and Li, 2005), housing prices (Kullman 2003), innovations in assorted state variables (Campbell and Vuolteenaho 2004; Brennan, Wang, and Xia, 2004; Petkova, 2005), liquidity risk (Pastor and Stambaugh, 2003; Acharya and Pedersen, 2005), and a host of conditioning variables that summarize the state of the economy, including the spread between low- and high-grade debt (Jagannathan and Wang, 1996), the aggregate consumption-to-wealth ratio (Lettau and Ludvigson, 2001), the housing collateral ratio (Lustig and Van Nieuwerburgh, 2004), the expenditure share of housing (Piazzesi, Schneider, and Tuzel, 2004), and the labor income to consumption ratio (Santos and Veronesi, 2005).

Empirically, many of the proposed models seem to do a good job explaining the size and B/M effects, an observation at once comforting and disconcerting: comforting because it suggests that rational explanations for the anomalies are readily available, disconcerting because it provides an embarrassment of riches. Reviewing the literature, one gets the uneasy feeling that it seems a bit too easy to explain the size and B/M effects. This is especially true given the great variety of factor models that seem to work, many of which have very little in common with each other.

Our paper is motivated by that suspicion. In particular, we explain why, despite the seemingly strong evidence that many proposed models can explain the size and B/M effects, we remain unconvinced by the evidence. We offer a critique of the empirical methods that have become popular in the asset-pricing literature, a number of prescriptions for improving the tests, and evidence that several of the proposed models don't work as well as originally advertised.

The heart of our critique is that the literature has often given itself an extremely low hurdle to meet in claiming success: high cross-sectional R^2s (or low pricing errors) when average returns on the Fama-French 25 size-B/M portfolios are regressed on their factor loadings. This hurdle is low because size and B/M portfolios are well-known to have a strong factor structure, i.e., the Fama and French's (1993) three factors explain more than 90% of the time-series variation in portfolios' returns and more than 75% of the cross-sectional variation in their average returns. Given those features, obtaining high cross-sectional R^2

is very easy because almost any proposed factor is likely to produce betas that line up with expected returns; just about all that's required is for a factor to be (weakly) correlated with SMB or HML but not with the tiny, idiosyncratic three-factor residuals of the size-B/M portfolios.

The problem we highlight is not just a sampling issue, i.e., it is not solved by getting standard errors right. In population, if returns have a covariance structure like that of size-B/M portfolios, true expected returns will line up with true factor loadings so long as a proposed factor is correlated with returns only through the two or three common sources of variation in returns. The problem is also not solved by using an SDF-GMM approach. Under the same conditions that give a high cross-sectional R^2, the true pricing errors in an SDF specification will be small or zero, a result that follows immediately from the close parallel between the regression and SDF approaches (see, e.g., Cochrane, 2001).

This is not to say that sampling issues aren't important. Indeed, the strong factor structure of the size-B/M portfolios also means that, even if we can find factors that have little ability to explain cross-sectional variation in the portfolios' true expected returns, we are still reasonably likely to find a high cross-sectional R^2 in sample. As an illustration, we simulate returns on artificial factors that, while correlated with returns, are constructed to have zero true cross-sectional R^2s for the size-B/M portfolios. We find that a sample adjusted R^2 might need to be as high as 50% to be statistically significant in models with one factor and as high as 75% in models with three factors. Further, with three factors, the power of the tests is essentially zero: the sampling distribution of the adj. R^2 is almost the same under the null of zero true R^2 as under alternatives as high as 90% true R^2. In short, the high R^2s reported in the literature aren't nearly as impressive as they might appear.

The obvious question then is: What can be done? What is the right way to test a proposed model given that standard tests aren't convincing? We offer three suggestions. First, since the problems are caused by the strong factor structure of size and B/M portfolios, one simple solution is to expand the set of test assets to include portfolios sorted in other ways, for example, by industry or factor loadings. Second, since the problems are exacerbated by the fact that empirical tests often ignore theoretical restrictions on the cross-sectional slopes, another simple solution is to take these restrictions seriously when theory provides appropriate guidance. For example, zero-beta rates should be quite close to the riskfree rate, the risk premium for a traded factor should be close to its average excess return, and the cross-sectional slopes in conditional models should be determined by the volatility of the conditional risk premium (as we explain later; see also Lewellen and Nagel, 2005). Last, since the problems are exacerbated by sampling issues, a third 'solution' is to report confidence intervals for the cross-sectional R^2 or squared pricing errors, something easily done via simulation.

We apply these prescriptions to a handful of proposed models from the recent literature. The results are disappointing. None of the five models that we consider performs well in our tests, despite the fact that all seemed quite promising in the original studies.

The paper proceeds as follows. Section 2 provides a skeptical view of typical asset-pricing tests, Section 3 offers our suggestions for improving the tests, and Section 4 applies these prescriptions to several recent models. Section 5 concludes.

2. A skeptical view of asset-pricing tests

Our analysis uses the following notation. Let R be the vector of excess returns on N test assets (in excess of the riskfree rate) and F be a vector of K risk factors that perfectly explain expected returns on the test assets, i.e., $\mu_R \equiv E[R]$ is linear in the $N \times K$ matrix of stocks' loadings on the factors, $B \equiv \text{cov}(R,F)\text{var}^{-1}(F)$. For simplicity, and without loss of generality, we assume that the mean of F equals the cross-sectional risk premium on B, implying $\mu_R = B\ \mu_F$. Thus, our basic model is

$$R = B\ F + e, \tag{1}$$

where e are mean-zero residuals with cov(e, F) = 0. We make no assumptions at this point about the covariance matrix of e, so the model is completely general (eq. 1 has no economic content).

We follow the convention that all vectors are column vectors unless otherwise noted. For generic random variables x and y, $\text{cov}(x, y) \equiv E[(x - \mu_x)(y - \mu_y)']$; i.e., the row dimension is determined by x and the column dimension is determined by y. We use ι to denote a conformable vector of ones, 0 to denote a conformable vector or matrix of zeros, and I to denote a conformable identity matrix. M denotes the matrix $I - \iota\iota'/d$ that transforms, through pre-multiplication, the columns of any matrix with row dimension d into deviations from the mean.

The factors in F can be thought of as a 'true' model that is known to price assets; it will serve as a benchmark but we won't be interested in it per se. Instead, we want to test a proposed model P consisting of J factors. The matrix of assets' factor loadings on P is denoted $C \equiv \text{cov}(R, P)\ \text{var}^{-1}(P)$, and we'll say that P 'explains the cross section of expected returns' if $\mu = C\ \gamma$ for some risk premium vector γ. Ideally,

γ would be determined by theory.

A common way to test whether P is a good model is to estimate a cross-sectional regression of expected returns on factor loadings

$$\mu = z\,\iota + C\,\lambda + \eta, \tag{2}$$

where λ denotes a $J \times 1$ vector of regression slopes. In principle, we could test three features of eq. (2): (i) z should be roughly zero (that is, the zero-beta rate should be close to the riskfree rate); (ii) λ should be non-zero and may be restricted by theory; and (iii) η should be zero and the cross-sectional R^2 should be one. In practice, empirical tests often focus only on the restrictions that $\lambda \neq 0$ and the cross-sectional R^2 is one (the latter is sometimes treated only informally). The following observations consider the conditions under which P will appear well-specified in such tests.

Observation 1. *Suppose F and P have the same number of factors and P is correlated with R only through the common variation captured by F, by which we mean that cov(e, P) = 0 (e is the residual in eq. 1). Assume, also, that the correlation matrix between F and P is nonsingular. Then expected returns are exactly linear in stocks' loadings on P – even if P has arbitrarily small (non-zero) correlation with F and explains very little of the time-series variation in returns.*

Proof: The assumption that cov(e, P) = 0 implies cov(R, P) = B cov(F, P). Thus, stocks' loadings on P are linearly related to their loadings on F: $C \equiv \text{cov}(R, P)\,\text{var}^{-1}(P) = B\,Q$, where $Q \equiv \text{cov}(F, P)\,\text{var}^{-1}(P)$ is the nonsingular matrix of slope coefficients when F is regressed on P. It follows that $\mu = B\,\mu_F = C\,\lambda$, where $\lambda = Q^{-1}\,\mu_F$. □

Observation 1 says that, if P has the same number of factors as F, testing whether expected returns are linear in betas with respect to P is essentially the same as testing whether P is uncorrelated with e – a test that doesn't seem to have much economic meaning in recent empirical applications. For example, in tests with size and B/M portfolios, we know that R_M, SMB, and HML (the 'true' model F in our notation) capture nearly all (more than 92%) of the time-series variation in returns, so the residual in R = B F + e is both small and largely idiosyncratic. In that setting, we don't find it surprising that almost any proposed macroeconomic factor P is correlated with returns primarily through R_M, SMB, and HML – indeed, we would be more surprised if cov(e, P) wasn't close to zero. In turn, we are not at all surprised that many proposed models seem to 'explain' the cross-section of expected size and B/M returns. The strong factor structure of size and B/M portfolios makes it likely that stocks' betas on almost any proposed factor will line up with their expected returns.[1]

Put differently, Observation 1 provides a skeptical interpretation of recent asset-pricing tests, in which unrestricted cross-sectional regressions (or equivalently SDF tests, as we explain below) have become the norm. In our view, the empirical tests say little more than that a number of proposed factors are correlated with SMB and HML, a fact that might have some economic content but seems like a pretty low hurdle to meet in claiming that a proposed model explains the size and B/M effects. We offer a number of suggestions for improving the tests below.

Observation 2. *Suppose returns have a strict factor structure with respect to F, i.e., var(e) is a diagonal matrix. Then any randomly chosen set of K assets perfectly explains the cross section of expected returns so long as the K assets aren't asked to price themselves (that is, the K assets aren't included as test assets on the left-hand side of the cross-sectional regression and the cross-sectional risk premia aren't required to equal the expected returns on the K assets). The only restriction is that R_K, the return on the K assets, must be correlated with F [cov(F, R_K) must be nonsingular].*

Proof: Let $P = R_K$ in Observation 1 and re-define R as the vector of returns for the remaining N – K assets and e as the residuals for these assets. The strict factor structure implies that $\text{cov}(e, R_k) = \text{cov}(e, B_K F + e_K) = 0$. The result then follows immediately from Observation 1. □

Observation 2 is useful for a couple of reasons. First, it provides a simple illustration of our argument that, in some situations, it is easy to find factors that explain the cross section of expected returns: under the fairly common assumption of a strict factor structure, any collection of K assets will work. Obtaining a high cross-sectional R^2 just isn't very difficult when returns have a strong factor structure, as they do in most empirical applications.

Second, Observation 2 illustrates that it can be important to take seriously restrictions on the cross-sectional slopes. In particular, Observation 2 hinges on the fact that the K asset factors aren't asked to price themselves, i.e., that the cross-sectional risk premia aren't restricted to equal the vector of expected

[1] The argument here works cleanly if a proposed model has at least three factors, which is often the case. The argument should also work fairly well even if P has only two factors since size and B/M portfolios all have multiple regression market betas close to one; in essence, the two-factor model of SMB and HML alone explains most of the cross-sectional variation of expected returns, so proposed models only need two factors (as long as we ignore any restriction on the intercept z).

returns on the K assets, as asset-pricing theory would predict. To see this, Observation 1 (proof) shows that the cross-sectional slopes on C are $\lambda = Q^{-1}\ \mu_F$, where Q is the matrix of slope coefficients when F is regressed on R_K. In the simplest case with one factor, λ simplifies to μ_K / ρ^2, where ρ is the correlation between R_K and F.[2] The slope λ is clearly greater than μ_K unless R_K is perfectly correlated with F. The implication is that the problem highlighted by Observations 1 and 2 – that 'too many' proposed factors explain the cross section of expected returns – would be less severe if the restriction on λ was taken seriously (e.g., R_K would then price the cross section only if $\rho = 1$).

Observations 1 and 2 are rather special since, in order to get clean predictions, we've assumed that a proposed model P has the same number of factors as the known model F. The intuition clearly goes through when J < K because, even in that case, we would expect the loadings on proposed factors to line up (imperfectly) with expected returns if the assets have a strong factor structure. The next observation generalizes our results, at the cost of changing the definitive conclusion in Observations 1 and 2 into a probabilistic statement.

Observation 3. *Suppose F has K factors and P has J factors, with $J \leq K$. Assume, as before, that P is correlated with R only through the factor F [cov(e, P) = 0], and that P and F are correlated, so that cov(F, P) has rank J. In a generic sense, made precise below, the cross-sectional R^2 in a regression of μ on C is expected to be J / K.*

Proof: By a 'generic sense,' we mean that we don't have any information about the contribution of each of the factors in F in explaining the cross-section of expected returns, and so treat the contributions as random. More specifically, suppose that the factor loadings on F satisfy $V_B = B'MB / N = I_K$, i.e., they are cross-sectionally uncorrelated and have unit variances; this assumption is without loss of generality since F can always be transformed to make the assumption hold. A 'generic sense' means that we view the risk premia on the transformed factors as being random draws from a normal distribution with mean zero and variance σ_γ^2. The proof then proceeds as follows: In a regression of μ on C, the risk premia are $\lambda = (C'MC)^{-1}C'M\mu$ and the R^2 is $\lambda'C'MC\lambda / \mu'M\mu$. By assumption $\mu = B\ \mu_F$ and Observation 1 (proof) shows that C = B Q, where $Q \equiv \text{cov}(F, P)\ \text{var}^{-1}(P)$. Substituting for λ, μ, and C, and using the assumption that $V_B = I_N$, the R^2 simplifies to $\mu_F'Q(Q'Q)^{-1}Q'\mu_F / \mu_F'\mu_F$, where $Q(Q'Q)^{-1}Q'$ is a symmetric, idempotent matrix of rank J. The risk premia, μ_F, are assumed MVN[0, σ_γ^2 I_K], as explained above, from which it

[2] $Q^{-1} = \text{var}(R_K) / \text{cov}(R_K, F)$ and, since F prices all the assets, $\mu_K = B_K\ \mu_F$, implying that $\mu_F = \mu_K\ B_K^{-1} = \mu_K\ \text{var}(F) / \text{cov}(R_K, F)$. Combining these equations gives the result.

follows that the R^2 has a Beta distribution with mean J / K. [The distribution of the R^2 follows from the fact that it can be expressed as a $z_1 / (z_1 + z_2)$, where z_1 and z_2 are independent, chi-squared variables with degrees of freedom J and K–J; see Muirhead, 1982, Thm. 1.5.7.] □

Observation 3 generalizes Observations 1 and 2. Our earlier results show that, if a K-factor model explains both the cross section of expected returns and much of the time-series variation in returns, then it should be easy to find other K-factor models that also explain the cross section of expected returns. The issue is a bit messier with J < K. Intuitively, the more factors that are in the proposed model, the easier it should be to find a high cross-sectional R^2 as long at the proposed factors are correlated with the 'true' factors. Thus, we aren't surprised at all if a proposed three-factor model explains the size and B/M effects about as well as the FF factors, nor are we surprised if a one- or two-factor model has some explanatory power. We *are* surprised if a one-factor model works as well as the FF factors, a result that is more impressive because it requires a single factor to capture the pricing information in both SMB and HML. [We note again that size–B/M portfolios all have Fama-French three-factor market betas close to one, so the model can be thought of as a two-factor model (SMB and HML) for the purposes of explaining cross sectional variation in expected returns.]

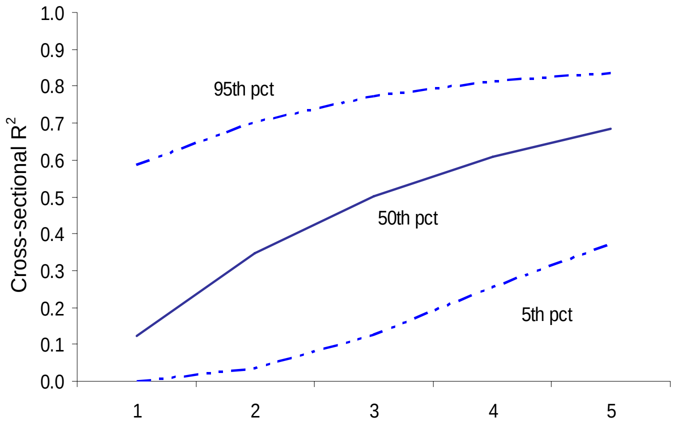
Figure 1 illustrates these results using Fama and French's 25 size-B/M portfolios, getting away from the specific assumptions underlying Observations 1 – 3. We calculate quarterly excess returns on the 25 portfolios from 1963 – 2004 and explore, in several simple ways, how easy it is to find factors that 'explain' the cross section of average returns. The figure treats the average returns and sample covariance matrix as population parameters; thus, like Observations 1 – 3, it focuses on explaining expected returns in population, not on sampling issues (which we consider later).

Each of the panels reports simulations using artificial factors to explain expected returns. In Panel A, the factors are constructed to produce a random vector of return betas: a 25 × 1 vector of loadings (for the 25 size-B/M portfolios) is randomly drawn from a MVN distribution with mean zero and covariance matrix proportional to the return covariance matrix. Thus, although the artificial factors aren't designed to explain expected returns, the loadings will tend to line up with expected returns (positively or negatively) simply because their cross-sectional pattern is determined by the covariance structure of returns. This procedure matches the spirit of Observations 1 – 3 but doesn't impose the requirement that the artificial factors covary only with common components in size-B/M portfolios (cov(e, P) doesn't have to be zero), though the common components will tend to dominate simply because they are so important.

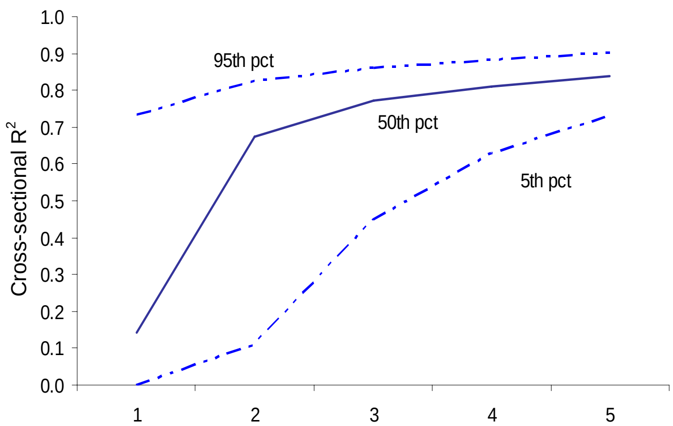
Figure 1. Population R^2s for artificial factors.

This figure explores how easy it is to find factors that explain, in population, the cross section of expected returns on Fama and French's 25 size-B/M portfolios. We randomly generate factors – either factor loadings directly or zero-investment factor portfolios, as described in the figure – and estimate the population R^2 when the size-B/M portfolios' expected returns are regressed on their factor loadings. The average returns and covariance matrix of the portfolios, quarterly from 1963 – 2004, are treated as population parameters in the simulations. The plots are based on 2,000 draws of 1 to 5 factors.

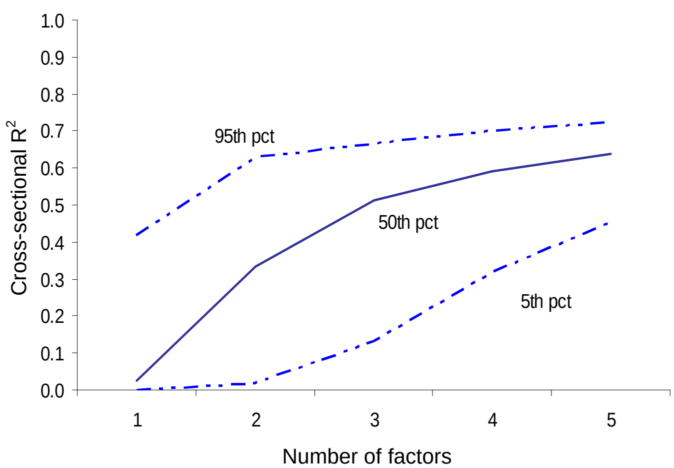
Panel A: Random draws of factor loadings.
Loadings for the 25 size-B/M portfolios are drawn from a MVN distribution with mean zero and covariance matrix proportional to the return covariance matrix.



Panel B: Random draws of factor portfolios.
Zero-investment factors, formed from the size-B/M portfolios, are generated by randomly drawing a 25×1 vector of weights from a standard normal distribution.



Panel C: Random draws of zero-mean factor portfolios.
Zero-investment factors, formed from the size-B/M portfolios, are generated by randomly drawing a 25×1 vector of weights from a standard normal distribution, but only factors with roughly zero expected returns are kept.



[An alternative interpretation of these simulations is to note that if we generate a time series of random factors uncorrelated with returns, the covariance matrix of *estimated* betas (in a multivariate regression of portfolio returns on the factor) is proportional to the covariance matrix of returns. Thus, the loadings in the simulations can be interpreted as sample betas for random ('useless') factors, and the population R^2 can be interpreted as a sample R^2 when size-B/M portfolios' average returns are regressed on these sample betas. The simulations show how often we expect to find high R^2s if researchers simply come up with factors that have nothing to do with returns.]

Panel A shows that it is easy to find factors that help explain expected returns on the size-B/M portfolios. With one factor, half of our factors produce an R^2 greater than 0.12 and 25% produce an R^2 greater than 0.30 (the latter isn't reported in the figure). With three factors, the median R^2 is 0.51 and the 75th percentile is 0.64, and with five factors, the median and 75th percentiles are a remarkable 0.68 and 0.76, respectively. Roughly half of our random three-factor models and 86% of our random five-factor models explain at least half of the cross-sectional variation in expected returns.

Panel B performs a similar exercise but, rather than randomly generate loadings, we randomly generate factors that are zero-investment combinations of the size-B/M portfolios: a 25×1 vector of weights is randomly drawn from a standard normal distribution with mean zero and variance one (the weights are then shifted and re-scaled to have a cross-sectional mean that is exactly zero and to have one dollar long and one dollar short). These simulations show how easy it is to stumble across factors that help explain the cross section of expected returns. As in Panel A, betas on the random factors will tend to line up with expected returns simply because of the covariance structure of returns, even though the factors aren't chosen to have any explanatory power. In fact, Panel B shows that the random factors here are even better at explaining expected returns: with one, three, and five factors, the median R^2s are 0.14, 0.77, and 0.84, while the 75th percentiles are 0.40, 0.81, and 0.87, respectively.

Finally, Panel C repeats the simulations in Panel B with a small twist: we keep only those random factors that have roughly zero expected returns [the factors in Panel B are expected to have zero expected returns ($E[\mu' x] = 0$ across draws of x) but don't because of random variation in x]. These simulations illustrate that it can be very important to impose restrictions on the cross-sectional slopes when possible; in particular, theory says that the risk premia on our random factors should be zero, equal to their expected returns, but Panel C ignores this restriction and just searches for the best possible fit in the cross-sectional regression. Thus, the actual R^2 differs from zero simply because we ignore the theoretical restrictions on the cross-sectional slopes and intercept. The additional degrees of freedom turn out to be very important, especially with multiple factors: with one, three, and five factors in Panel C, the median R^2s are 0.03, 0.51, and 0.64, while the 75th percentiles are 0.12, 0.60, and 0.68, respectively (again, properly restricted R^2s would all be close to zero).

The results above illustrate that the factor structure of size-B/M portfolios makes it easy to find factors that produce high *population* cross-sectional R^2s. Our final two observations show that the problem is similar in SDF–GMM tests and exacerbated by sampling issues.

Observation 4. *Suppose F has K factors and P has J factors, with $J \leq K$. Assume, as before, that P is correlated with R only through the factor F [cov(e, P) = 0], and that P and F are correlated, so that cov(F, P) has rank J. In a generic sense, made precise below, the sum of squared pricing errors in an SDF framework (i.e., $\varepsilon = E[mR]$) are expected to be q (K – J), where q is defined below. The pricing errors are exactly zero when J = K.*

Proof: By a 'generic sense,' we again mean that we don't have any information about the contribution of each factor in F in explaining the cross section of expected returns, but we operationalize the idea slightly differently than in Observation 3 (a similar result holds if we use the earlier definition but the proof is messier). Specifically, suppose the factor loadings on F satisfy $B'B / N = I_K$; this assumption is without loss of generality since F can always be transformed to make the assumption hold. A 'generic sense' means that we view the risk premia on the transformed B as being random IID draws from a normal distribution with mean α and variance σ_γ^2. The proof proceeds as follows: Define the SDF as $m = a - b'P$ and the pricing errors as $\varepsilon = E[mR]$. Using excess returns, the SDF is defined up to a constant of proportionality (see Cochrane, 2001); we assume that P has mean zero and chose $a = 1/(1+r_f)$ in order to impose the constraint that $E[m] = 1/(1+r_f)$. In first-stage GMM, $\min_b \varepsilon'\varepsilon$, the solution is $b = a(D'D)^{-1}D'\mu$ and $\varepsilon = a[I_N - D(D'D)^{-1}D']\mu$, that is, GMM is equivalent to a cross-sectional regression of μ on $D \equiv \text{cov}(R, P)$. The sum of squared pricing errors is $\varepsilon'\varepsilon = a^2\mu'[I_N - D(D'D)^{-1}D']\mu$. By assumption $\mu = B\,\mu_F$ and $B'B/N = I_K$, so $\varepsilon'\varepsilon$ can be re-written as $a^2N\mu_F'[I_K - B'D(D'D)^{-1}D'B/N]\mu_F$. The matrix $H = I_K - B'D(D'D)^{-1}D'B/N$ is symmetric and idempotent with rank K – J [this uses the fact that $B'B/N = I_K$, $C = D\,\text{var}^{-1}(P)$, and $C = B\,Q$, where C is the factor loadings on P and $Q \equiv \text{cov}(F, P)\,\text{var}^{-1}(P)$]. The risk premia are assumed $\text{MVN}[\alpha, \sigma_\gamma^2\, I_K]$, as explained above, from which it follows that $\mu_F'H\mu_F$ is proportional to a non-central chi-squared variate, with mean $(\alpha^2 + \sigma_\gamma^2)\,(K - J)$ (ε and $\mu_F'H\mu_F$ are exactly zero when J = K). The sum of squared errors has expectation $E[\varepsilon'\varepsilon] = a^2\,N\,(\alpha^2 + \sigma_\gamma^2)\,(K - J)$. □

Observation 4 is the SDF equivalent of our earlier cross-sectional R^2 results. It says that, as long as proposed factors P covary with returns only through the factors F – an assumption that seems likely to hold for just about any proposed factor when the test assets are size-B/M portfolios – the model will help reduce the pricing errors $\varepsilon = E[mR]$. The errors are (expected to be) smaller the more factors that are in P and, in the limit, drop to zero when P has the same number of factors as F. The magnitude of the errors when $J < K$ can be interpreted by noting that with no factor, $J = 0$, the sum of squared pricing errors is expected to be $a^2 N (\alpha^2 + \sigma_\gamma^2) K$ given our assumptions. Thus, every factor reduces $E[\varepsilon'\varepsilon]$ by a faction 1/K. The decline in pricing errors is nearly mechanical in tests with size-B/M portfolios, because of their strong factor structure, and has little economic meaning.

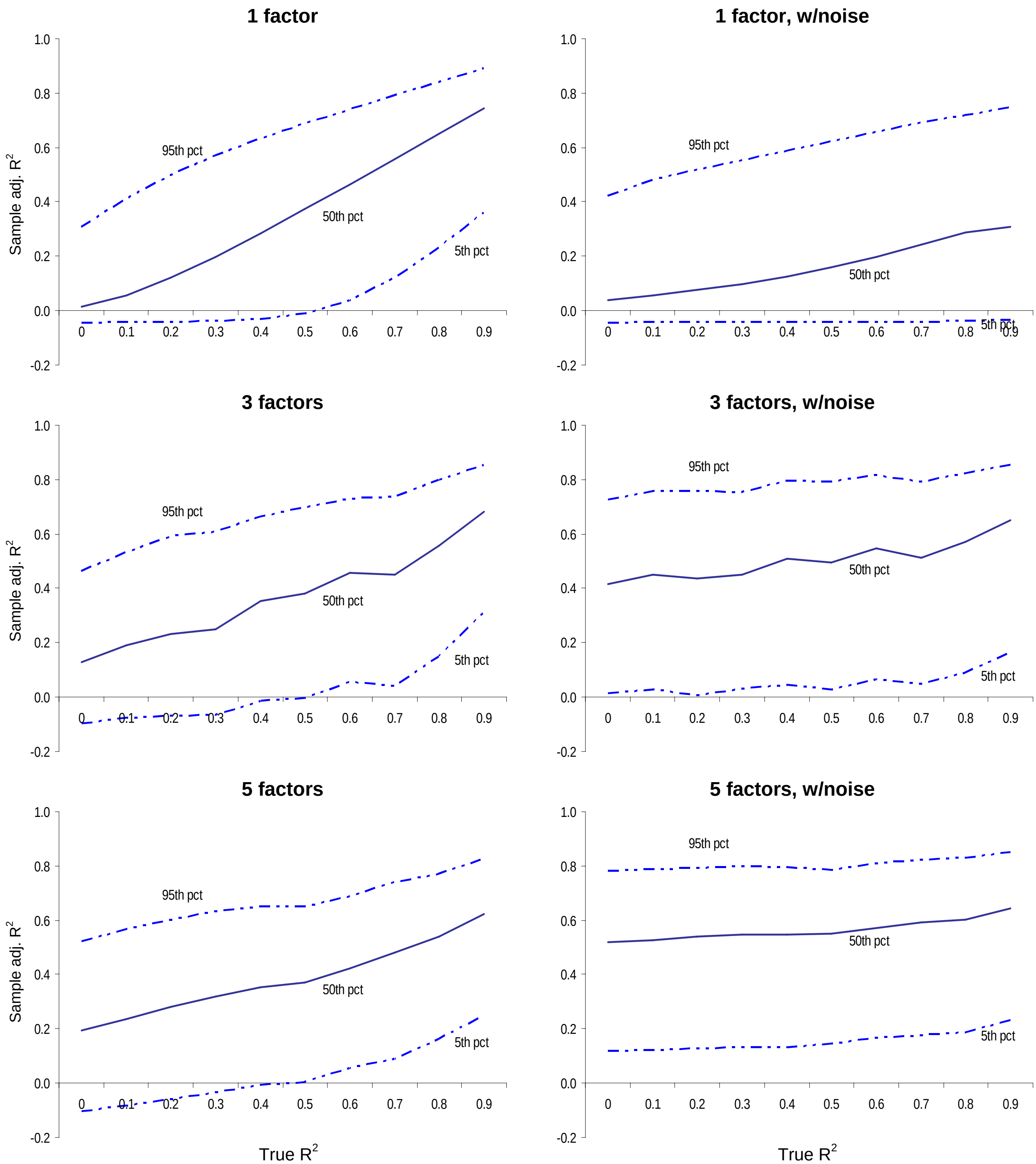
Observation 5. *The problems are exacerbated by sampling issues: If returns have a strong factor structure, it can be easy to find a high sample cross-sectional R^2 even in the unlikely scenario that the population R^2 is small or zero.*

Observation 5 is intentionally informal and, in lieu of a proof, we offer simulations using Fama and French's 25 size-B/M portfolios to illustrate the point. The simulations are different from Figure 1 because, rather than study the population cross-sectional R^2 for artificial factors, we now focus on sampling variation in estimated R^2s *conditional on a given population R^2*. The simulations have two steps: First, we fix a true cross-sectional R^2 that we want a proposed model to have and randomly generate a vector of factor loadings C that produces that R^2 (approximately). A factor portfolio $P = w'R$ is constructed that has those factor loadings, i.e., we find the vector of portfolio weights w such that $\text{cov}(R, P) = \text{var}(R)\, w$ is proportional to C. Second, we bootstrap artificial time series of returns and factors by sampling, with replacement, from the historical time series of size-B/M returns (quarterly returns from 1963 – 2004). We then estimate the sample cross-sectional adj. R^2 for the artificial data by regressing average returns on estimated factor loadings. The second step is repeated 2,000 times to construct a sampling distribution of the adj. R^2. In addition, to make sure the particular vector of loadings generated in step 1 isn't important, we repeat that step 10 times, giving us a total sample of 20,000 estimated adjusted R^2s corresponding to an assumed true R^2.

Figure 2 shows results for models with 1, 3, and 5 factors. The left-hand column plots the distribution of the sample adjusted R^2 (5th, 50th, and 95th percentiles) corresponding to true R^2s of 0.00 to 0.90 for models in which the factors are portfolio returns, as described above. The right-hand column repeats the exercise but uses factors that are only imperfectly correlated with returns, as they are in most empirical applications; we start with the portfolio factors used in the left-hand panels and add noise equal to 2/3 of

Figure 2: Sample distribution of the cross-sectional adj. R^2.

This figure shows the sample distribution of the cross-sectional adj. R^2 (average returns regressed on estimated factor loadings) for Fama and French's 25 size-B/M portfolios from 1963 – 2004 (quarterly returns). The plots use one to five randomly generated factors that together have the true R^2 reported on the x-axis. In the left-hand panels, the factors are combinations of the size-B/M portfolios (the weights are randomly drawn to produce the given R^2, as described in the text). In the right-hand panels, noise is added to the factors equal to 2/3 of a factor's total variance, to simulated factors that are not perfectly spanned by returns. The plots are based on 20,000 bootstrap simulations (10 sets of random factors; 2,000 simulations with each).



their total variance. Thus, for the right-hand plots, a maximally correlated combination of the size-B/M portfolios would have a time-series R^2 of 0.33 with each factor.

The figure shows that a sample R^2 needs to be quite high to be statistically significant, especially for models with several factors. Focusing on the right-hand column, the 95th percentile of the sampling distribution using one factor is 42%, using three factors is 73%, and using five factors is 78% – when the true cross-sectional R^2 is zero! Thus, even if we could find factors that have no true explanatory power (something that seems unlikely given our population results above), it still wouldn't be terribly surprising to find fairly high R^2s in sample. Further, with either three or five factors, the ability of the sample R^2 to discriminate between good and bad models is essentially nil, since the distribution of the sample R^2 is almost identical when the true R^2 is 0.0 or 0.90. For example, with five factors, a sample R^2 greater than 80% is needed to reject that the true R^2 is 30% or less (at a 5% significance level), but that outcome is unlikely even if the true R^2 is 0.90 (probability of only 0.14). The bottom line is that, in both population and sample, a high cross-sectional R^2 seems to provide very little information about whether a proposed model is good or bad.

Related research

Our appraisal of asset-pricing tests overlaps with a number of studies. Roll and Ross (1994) and Kandel and Stambaugh (1995) argue that the cross-sectional R^2 in simple CAPM tests isn't very meaningful because, as a theoretical matter, it tells us little about the location of the market proxy in mean-variance space (see also Kimmel, 2003). We reach a similarly skeptical conclusion about the R^2 but emphasize different issues. The closest overlap comes from our simulations in Panel C of Figure 1, which show that factor portfolios with zero mean returns might still produce high R^2 in unrestricted cross-sectional regressions. These portfolios, by construction, are far from the mean-variance frontier – they have zero Sharpe ratios – yet often have high explanatory power, consistent with the results of Roll and Ross and Kandel and Stambaugh.

Kan and Zhang (1999) study cross-sectional tests with 'useless' factors, defined as a factor that is uncorrelated in population with returns. They show that the usual asymptotics break down because the cross-sectional spread in estimated loadings goes to zero as T gets big (since all the loadings go to zero). Our simulations in Panel A of Figure 1 have some overlap since, as pointed out earlier, they can be interpreted as showing the sample R^2 when randomly generated useless factors are used to explain returns. The issues are different since our simulations generate random factors but hold the time series of returns constant (thus, they don't really consider the sampling issues discussed by Kan and Zhang). More broadly, our results are different because we focus on population R^2s and, when we do look at sampling distributions in Fig. 2, the factors are not 'useless.'

Some of our results are reminiscent of the literature on testing the APT and multifactor models (see, e.g., Shanken 1987, Reisman, 1992; Shanken, 1992a). Most closely, Nawalkha (1997) derives results like Observations 1 and 2 above, though his focus is different. In particular, he emphasizes that, in the APT, 'well-diversified' variables (those uncorrelated with idiosyncratic risks) can be used in place of the 'true' factors without any loss of pricing accuracy. We generalize his theoretical results to models with $J < K$ proposed factors, consider sampling issues, and emphasize the empirical implications for recent tests using size-B/M portfolios.

Finally, our critique is similar in spirit to a contemporaneous paper by Daniel and Titman (2005). They show that, even if characteristics determine expected returns (e.g., expected returns are linear in B/M), a proposed factor can appear to price characteristic-sorted portfolios simply because, in the underlying population of stocks, factor loadings and characteristics are correlated (forming portfolios tends to inflate that correlation). Our ultimate conclusions about using characteristic-sorted portfolios are similar but we highlight different concerns, emphasizing the importance of the factor structure of size-B/M portfolios, the impact of using many factors and not imposing restrictions on the cross-sectional slopes, and the role of both population and sampling issues.

3. How can we improve empirical tests?

The theme of Observations 1 – 5 is that, in situations like those encountered in practice, it may be easy to find factors that 'explain' the cross section of expected returns. A high cross-sectional R^2 (or small pricing errors in SDF tests) often has little economic meaning and, in our view, should not be taken as providing much support for a proposed model. The problem is not just a sampling issue – it cannot be solved by getting standard errors right – though sampling issues exacerbate the problem. Here, we offer a few suggestions for improving empirical tests.

Prescription 1. Expand the set of test portfolios beyond size–B/M portfolios.

Given the importance of the size and value anomalies, empirical tests often focus exclusively on size-B/M portfolios. This practice is understandable but problematic, since the concerns highlighted above are most severe when a couple of factors explain nearly all of the time-series variation in returns, as is true for size-B/M portfolios. One simple solution, then, is to include portfolios that don't correlate as strongly with

SMB and HML. Reasonable choices include industry–, beta–, volatility–, or factor-loading–sorted portfolios (the last being loadings on a proposed factor; an alternative would be to use individual stocks in the regression, though errors-in-variables problems could make this impractical). Bond portfolios might also be used. The idea is to price the portfolios all at the same time, not in separate cross-sectional regressions. Also, the additional portfolios don't need to offer a big spread in expected returns; the goal is simply to relax the tight factor structure of size-B/M portfolios.

Figure 3 illustrates this idea. We replicate the simulations in Figure 1 but, rather than focus solely on size-B/M portfolios, we augment them with Fama and French's 30 industry portfolios. As before, we generate artificial factors and explore how well they explain, in population, the cross section of expected returns (the portfolios' average returns and covariances from 1963 – 2004 are treated as population parameters). The factors are generated in three ways. In Panel A, the factors are constructed to produce a randomly chosen 55×1 vector of factor loadings, drawing from a MVN distribution with mean zero and covariance matrix proportional to the return covariance matrix. In Panel B, the factors are constructed by randomly drawing a 55×1 vector of portfolio weights from a standard normal distribution. And in Panel C, we repeat the simulations of Panel B by keep only the factor portfolios that have (roughly) zero expected returns. The point in each case is to explore how easy it is to find factors that produce high cross-sectional R^2 (in population). We refer the reader to the discussion of Figure 1 for the logic and interpretation of each set of simulations.

Figure 3 shows that it is much 'harder' – using artificial factors – to explain expected returns for the 55 portfolios than for the 25 size-B/M portfolios (the median and 95th percentiles for the latter are repeated from Fig. 1 for comparison). For example, with three factors, the median R^2 for the full set of 55 portfolios is 15% in Panel A, 20% in Panel B, and 10% in Panel C, compared with median R^2s for the 25 size-B/M portfolios of 50%, 77%, and 51%, respectively. The difference between the 25 size-B/M portfolios and the full set of portfolios is largest for models with at least three factors, consistent with the three-factor structure of size-B/M portfolios being important. In short, the full set of portfolios seems to provide a more rigorous test of a proposed model.

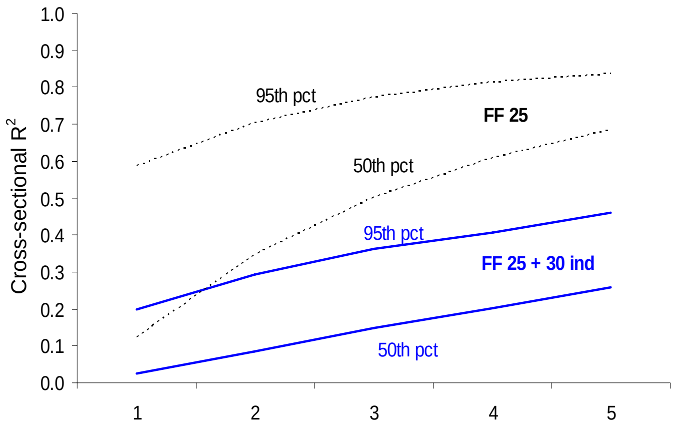
Prescription 2. Take the magnitude of the cross-sectional slopes seriously.

Too often in the recent literature, papers focus on a model's cross-sectional R^2 and don't bother to check whether the estimated slopes and zero-beta rates are reasonable. Yet theory often provides guidance for both that should be taken seriously, i.e., the theoretical restrictions should be imposed ex ante or tested ex post. Most clearly, theory says the zero-beta rate should equal the riskfree rate. The standard retort is that

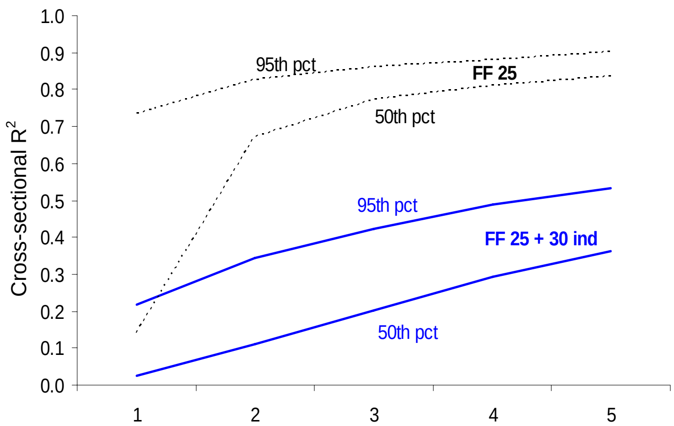
Figure 3. Population R^2s for randomly generated factors: Size-B/M and industry portfolios.

This figure compares how easy it is to find factors that explain, in population, the cross section of expected returns on Fama and French's 25 size-B/M portfolios (dotted lines) vs. 55 portfolios consisting of the 25 size-B/M portfolios and Fama and French's 30 industry portfolios (solid lines). We randomly generate factors – either factor loadings directly or zero-investment factor portfolios, as described in the figure – and estimate the population R^2 when the portfolios' expected returns are regressed on their factor loadings. The average returns and covariance matrix of the portfolios, quarterly from 1963 – 2004, are treated as population parameters in the simulations. The plots are based on 2,000 draws of 1 to 5 factors.

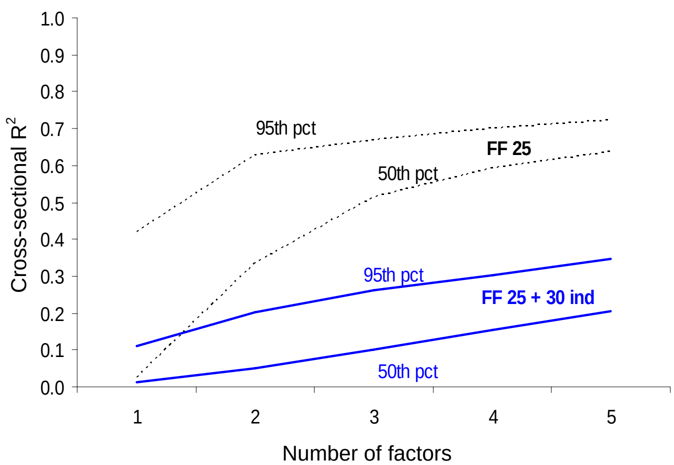
Panel A: Random draws of factor loadings.
Loadings for the 25 size-B/M portfolios are drawn from a MVN distribution with mean zero and covariance matrix proportional to the return covariance matrix.



Panel B: Random draws of factor portfolios.
Zero-investment factors, formed from the size-B/M portfolios, are generated by randomly drawing a 25×1 vector of weights from a standard normal distribution.



Panel C: Random draws of zero-mean factor portfolios.
Zero-investment factors, formed from the size-B/M portfolios, are generated by randomly drawing a 25×1 vector of weights from a standard normal distribution, but only factors with roughly zero expected returns are kept.



Black's (1972) model relaxes this constraint if borrowing and lending rates differ, but this argument isn't convincing in our view: borrowing and lending rates just aren't sufficiently different (perhaps 1% annually) to justify the extremely high zero-beta estimates in many papers. An alternative argument is that the equity premium is anomalously high, à la Mehra and Prescott (1985), so it's unreasonable to ask a consumption-based model to explain it. But its not clear why we should view a model favorably if it cannot explain the level of expected returns. Would researchers be satisfied with the standard consumption CAPM if it explained cross-sectional variation in returns but got the equity premium wrong by an order of magnitude? We doubt it, and for the same reason we would reject any model that gives an excessive zero-beta rate.

A related restriction, mentioned earlier, is that the risk premium for any factor portfolio should be the portfolio's expected excess return. For example, the cross-sectional price of market-beta risk should equal the equity premium; the price of yield-spread risk, captured by movements in long-term Tbond returns, should be the expected Tbond return over the riskfree rate. In practice, this type of restriction could be tested in cross-sectional regressions or, better yet, imposed ex ante by focusing on time-series regression intercepts (Jensen's alphas). Shanken and Weinstein (2005) show how to impose this restriction when only some of factors are returns. Below, we suggest a couple of simple ways to incorporate the constraint into cross-sectional regressions.

As a third example, conditional asset-pricing models often imply concrete restrictions on cross-sectional slopes, a point emphasized by Lewellen and Nagel (2005). For example, as Jagannathan and Wang (1996) show, a one-factor conditional CAPM implies a two-factor unconditional model: $E_{t-1}[R_t] = \beta_t \gamma_t \rightarrow E[R] = \beta \gamma + \text{cov}(\beta_t, \gamma_t)$, where β_t and γ_t are the conditional beta and equity premium, respectively, and β and γ are their unconditional means. Recent studies explore the two-factor cross-sectional regression $E[R_i] = \beta_i \gamma + \text{cov}(\beta_{it}, \gamma_t)$ using various estimates of stocks' betas and 'beta-premium sensitivities' $\varphi_i = \text{cov}(\beta_{it}, \gamma_t)$. The slope on φ_i should be one in this regression yet that constraint is often ignored. [Put differently, the CAPM has at most one free parameter in the cross section; studies that freely estimate two or three cross-sectional slopes must, by logical deduction, ignore some constraints.] Lewellen and Nagel discuss these issues in detail and provide empirical examples from recent tests of both the simple and consumption CAPMs. For tests of the simple CAPM, the constraint is easily imposed using the conditional time-series regressions of Shanken (1990), if the relevant state variables are all known, or the short-window approach of Lewellen and Nagel, if they are not.

Prescription 3. If a proposed factor is a traded portfolio, include the factor as one of the test assets on the left-hand side of the cross-sectional regression.

Prescription 3 builds on Prescription 2, in particular, the idea that the cross-sectional price of risk for a traded factor should be the factor's expected excess return. One simple way build this restriction into a cross-sectional regression is to ask the factor to price itself; that is, to test whether the factor itself lies on the estimated cross-sectional regression line.

An interesting case when Prescription 3 is most important is when the cross-sectional regression is estimated with GLS rather than OLS. In particular, it can be shown that, if a traded factor is included as a left-hand side portfolio, GLS forces the regression to price the asset perfectly: the estimated slope on the traded factor's loading exactly equals the factor's average return in excess of the zero-beta rate (in essence, the asset is given infinite weight in the regression). Thus, a GLS cross-sectional regression, when the traded factor is included as a test portfolio, is similar to the time-series approach of Black, Jensen, and Scholes (1972) and Gibbons, Ross, and Shanken (1989).

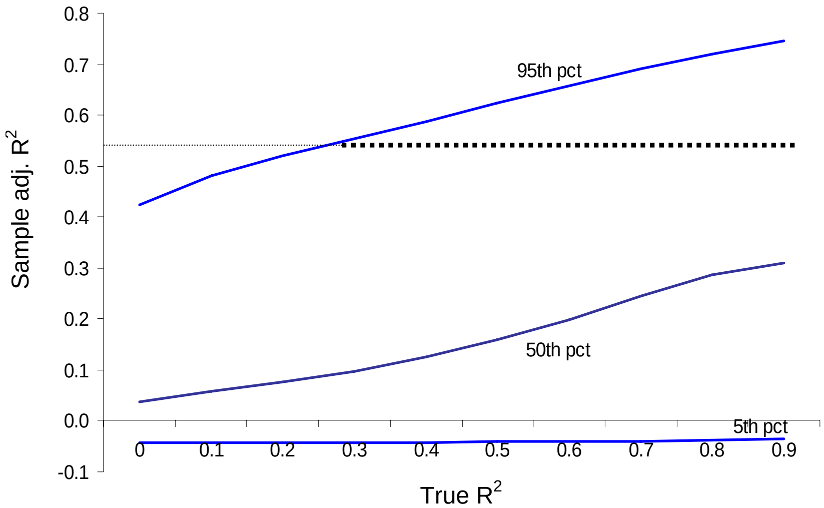
Prescription 4. Report confidence intervals for the cross-sectional R^2.

Prescription 4 is less a solution to the problems highlighted above – indeed, it does nothing to address the concern that it may be easy to find factors that produce high population R^2s – than a way to make the sampling issues more transparent. We suspect researchers would put less weight on the cross-sectional R^2 if the extremely high sampling error in it was clear (extremely high when using size-B/M portfolios, though not necessarily with other assets). More generally, we find it odd that papers often emphasize this statistic without regard to its sampling properties.

The distribution of the sample R^2 can be derived analytically in special cases, but we're not aware of a general formula or one that incorporates first-stage estimation error in factor loadings. An alternative is to use simulations like those in Figure 2, one panel of which is repeated in Figure 4. The simulations indicate that the sample R^2 is often significantly biased and skewed by an amount that depends on the true cross-sectional R^2. These properties suggest that reporting a confidence interval for R^2 is more meaningful than reporting just a standard error.

The easiest way to get confidence intervals is to 'invert' Figure 4, an approach suggested by Stock (1991) in a different context. In the figure, the sample distribution of the estimated R^2, for a given true R^2, is found by slicing the picture along the x-axis (fixing x, then scanning up and down). Conversely, a

Figure 4. Sample distribution of the cross-sectional adj. R^2.
This figure repeats the '1 factor, w/noise' panel of Fig. 2. It shows the sample distribution of the cross-sectional adj. R^2, as a function of the true cross-sectional R^2, for a model with one factor using Fama and French's 25 size-B/M portfolios from 1963 – 2004 (quarterly returns). The simulated factor is a combination of the size-B/M portfolios (the weights are randomly drawn to produce the given R^2, as described in Section 2) plus noise equal to 2/3 of the factor's total variance. The plot is based on 20,000 bootstrap simulations (10 sets of simulated factors; 2,000 simulations with each).



confidence interval for the true R^2, given a sample R^2, is found by slicing the picture along the y-axis (fixing y, then scanning across). For example, a sample R^2 of 0.55 implies a 90% confidence interval for the population R^2 of roughly [0.30, 1.00], depicted by the dark dotted line in the graph. The confidence interval represents all values of the true R^2 for which the estimated R^2 falls within the 5th and 95th percentiles of the sample distribution. The extremely wide interval in this example illustrate just how uninformative the sample R^2 can be.

Prescription 5. Report confidence intervals for the (weighted) sum of squared pricing errors.

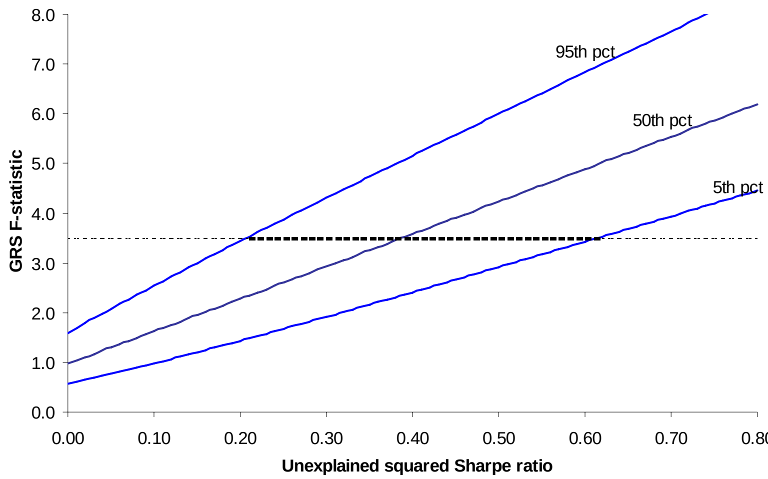
Prescription 5 has the same goal as Prescription 4, to provide a better summary measure of how well a model performs and to make sampling issues more transparent. Again, Prescription 5 does nothing to address our concern that it is easy to find factors that produce small population pricing errors for size-B/M portfolios. But reporting confidence intervals should at least make clear when a test has low power: we may not reject that a model performs perfectly (the null of zero pricing errors), but we also won't reject that the pricing errors are quite large. The opposite, of course, can be true as well: confidence intervals can reveal if a model is rejected not because the pricing errors are economically large but because the tests simply have strong power.

The (weighted) sum of squared pricing errors (SSPE) is an alternative to the cross-sectional R^2 as a measure of performance. Like the R^2, sample estimates of the SSPE are generally strongly biased and skewed, suggesting that confidence intervals are more meaningful than standard errors or p-values. The literature considers several versions of such statistics, including Shanken's (1985) cross-sectional T^2 statistic, Gibbons, Ross, and Shanken's (GRS 1989) F-statistic, Hansen's (1982) J-statistic, and Hansen and Jagannathan's (1997) HJ-distance. Confidence intervals for any of these could be obtained using an approach like Figure 4, plotting the sample distribution as a function of the true parameter. We describe here how to get confidence intervals for the GRS F-test and the closely related HJ-distance, both of which have useful economic interpretations and impose appropriate restrictions on the zero-beta rate and risk premia (assuming, in the case of the HJ-distance, that a riskless asset is included).

The GRS F-statistic tests whether the time-series intercepts (pricing errors) are all zero when excess returns are regressed on a set of factor portfolios, $R = a + C\ P + e$. (The F-test can be used only if the factors are all portfolio returns or if the non-return factors are replaced by mimicking portfolios.) Let $\hat{a}$ be the OLS estimates of a and let $\Sigma \equiv \mathrm{var}(e)$. The covariance matrix of $\hat{a}$, given a sample for T periods, is $\Omega = c\ \Sigma$, where $c = (1 + s_P^2) / T$ and s_P^2 is the sample maximum squared Sharpe ratio attainable from combinations of P. Gibbons et al. (1989) show that, under standard assumptions, the weighted sum of squared pricing errors, $S = \hat{a}'\ \Omega_{OLS}^{-1}\ \hat{a} = c^{-1}\ \hat{a}'\ \Sigma_{OLS}^{-1}\ \hat{a}$ is asymptotically chi-squared and, if returns are multivariate normal, the statistic $F = c^{-1}\ \hat{a}'\ \Sigma_{OLS}^{-1}\ \hat{a}\ [(T-N-K) / N\ (T-K-1)]$ is small-sample F with non-centrality parameter $\lambda = c^{-1}\ a'\ \Sigma^{-1}\ a$ and degrees of freedom N and $T - N - K$. Moreover, the quadratic $\theta_z^2 = a'\ \Sigma^{-1}\ a$ is the model's unexplained squared Sharpe ratio, the difference between the population squared Sharpe ratio of the tangency portfolio (θ_τ^2) and that attainable from P (θ_P^2). Thus, a confidence interval for θ_z^2 can be found by inverting a graph like Fig. 4, showing the sample distribution of F as a function of θ_z^2 (or, more formally, finding the set of θ_z^2 for which F is less than, say, the 95th percentile of an F-distribution with non-centrality parameter $c^{-1}\ \theta_z^2$).

The HJ-distance is similar but tests whether the SDF pricing errors, $\varepsilon = E[y(1+R) - 1]$, are zero, where $y = a + b\ P$ is a proposed SDF and we now define R as an N+1 vector of total (not excess) returns, including the riskless asset as a test asset. Let m be any well-specified SDF. Hansen and Jagannathan (1997) show that the distance between y and the set of true SDFs, $D = \min_m E[(y-m)^2]$, also equals the largest squared pricing error available on any portfolio x, relative to its second moment, i.e., $D = \max_x (\varepsilon' x)^2 / E[(1+R_x)^2]$. Using the second definition, the distance is easily shown to equal $D = \varepsilon'\ H^{-1}\ \varepsilon$, where $H \equiv E[(1+R)(1+R)']$ is the second moment matrix of gross returns. To get a confidence interval for D, the Appendix shows

Figure 5. Sample distribution of the GRS F-statistic and confidence interval for θ_z^2.
This figure provides a test of the CAPM using quarterly excess returns on Fama and French's 25 size-B/M portfolios from 1963–2004. The sample distribution of the GRS F-statistic, for a given value of the true unexplained squared Sharpe ratio, θ_z^2, can be found by slicing the graph along the x-axis (fixing θ_z^2, then scanning up and down to find the 5th, 50th, and 95th percentiles of the sample distribution). Alternatively, a confidence interval for θ_z^2, given the GRS F-statistic, is found by slicing the picture along the y-axis (fixing y, then scanning across). θ_z^2 is the difference between the squared Sharpe ratio of the tangency portfolio and that of the CRSP value-weighted index. GRS F = c^{-1} $\hat{a}' \Sigma_{OLS}^{-1} \hat{a}$ (T–N–1)/ [N(T–2)]; it has an F-distribution with non-centrality parameter c^{-1} θ_z^2 and degrees of freedom N and T–N–1, where N = 25, T = 168, and $c \approx 1/T$. The sample F-statistic is 3.491 and the corresponding 90% confidence interval for θ_z^2 is depicted by the dark dotted line.



that $D = \theta_z^2 / (1 + r_f)^2$, where θ_z^2 is defined in the previous paragraph. Thus, like the GRS F-statistic, the estimate of D is small-sample F up to a constant of proportionality.[3] A confidence interval is then easily obtained using the approach described above.

Figure 5 illustrates the confidence interval approach using the GRS F-statistic to test the unconditional CAPM. The tests use quarterly excess returns on Fama and French's 25 size-B/M portfolios from 1963 – 2004 and our market proxy is the CRSP value-weighted index. The size and B/M effects are quite strong during this sample (the average absolute quarterly alpha is 0.96% across the 25 portfolios), and the GRS F-statistic strongly rejects the CAPM, F = 3.491 with a p-value of 0.000. The graph shows, moveover, that we can reject that the squared Sharpe ratio on the market is within 0.21 of squared Sharpe ratio of the tangency portfolio: a 90% confidence interval for $\theta_z^2 = \theta_\tau^2 - \theta_M^2$ is [0.21, 0.61]. Interpreted differently,

[3] We assume that the factors in P are either portfolios or that all non-return factors have been replaced by mimicking portfolios. In the latter scenario, the small-sample F-distribution would not take into account estimation error in the mimicking-portfolio weights, though it may provide a reasonable approximation nonetheless (Shanken and Zhou, 2005, provide evidence that the F-distribution works well in a related context). Simulations could be used to obtain the sampling distribution in place of the assumed F distribution.

following MacKinlay (1995), there exists a portfolio z that is uncorrelated with the market and, with 90% confidence, has a quarterly Sharpe ratio between 0.46 (=$0.21^{1/2}$) and 0.78 (=$0.61^{1/2}$). The confidence interval provides a good summary measure of just how poorly the CAPM works.

4. Empirical examples

The prescriptions above are typically easy to implement and, while not a complete solution to the problems discussed in Section 2, should help to improve the power and rigor of empirical tests. As an illustration, we report tests for several models that have been proposed recently in the literature. Cross-sectional tests in the original studies focused on Fama and French's 25 size-B/M portfolios, precisely the scenario for which our concerns are greatest. Our goal here is not to disparage the papers – indeed, we believe the studies provide economically important insights – nor to provide a full review of the (often extensive) empirical tests in each paper, but only to show that our prescriptions can dramatically change how well a model seems to work.

We select models to test for which necessary data are readily available. The models include: (i) Lettau and Ludvigson's (2001) conditional consumption CAPM (CCAPM), in which the conditioning variable is the aggregate consumption-to-wealth ratio *cay* (available on Ludvigson's website); (ii) Lustig and Van Nieuwerburgh's (2004) conditional CCAPM, in which the conditioning variable is the housing collateral ratio *mymo* (we consider only their linear model with separable preferences; *mymo* is available on Van Nieuwerburgh's website;); (iii) Santos and Veronesi's (2004) conditional CAPM, in which the conditioning variable is the labor income-to-consumption ratio s^{ω}; (iv) Li, Vassalou, and Xing's (2005) investment model, in which the factors are investment growth rates for households (ΔI_{HH}), non-financial corporations (ΔI_{Corp}), and the non-corporate sector (ΔI_{Ncorp}) (we consider only this version of their model); and (v) Yogo's (2005) durable–consumption CAPM, in which the factors are the growth in durable and non-durable consumption, Δc_{Dur} and Δc_{Non}, and the market return (R_M) (we consider only his linear model; the consumption series are available on Yogo's website).

Table 1 reports cross-sectional regressions of average returns on estimated factor loadings for the five models. The tests use quarterly excess returns (in %), from 1963 – 2004, and highlight Prescriptions 1 and 4, our suggestions to expand the set of test portfolios beyond size-B/M portfolios and to report a confidence interval for the cross-sectional R^2. Specifically, we compare results using Fama and French's 25 size-B/M portfolios alone ('FF25' in the table) with results for the expanded set of 55 portfolios that includes Fama and French's 30 industry portfolios ('FF25 + 30 ind'). Our choice of industry portfolios is based on the notion that they should provide a fair test of the models (in contrast to, say, momentum portfolios whose returns seem to be anomalous). To get confidence intervals for the R^2, we use the method suggested in the last section: we simulate the distribution of the sample adjusted R^2 for true R^2s between 0.0 to 1.0 and invert plots like Figure 4. The simulations are similar to those in Figures 2 and 4, with the actual factors for each model now used in place of the artificial factors.[4]

Table 1 shows three key results. First, adding industry portfolios dramatically changes the performance of the models, in terms of slope estimates, cross-sectional R^2s, and χ^2 statistics (the latter tests whether pricing errors for a model are jointly zero, calculated as in Shanken, 1985). Compared with regressions using only size-B/M portfolios, the slope estimates are almost always closer to zero and the cross-sectional R^2s often drop substantially. The adj. R^2 drops from 58% to 0% for Lettau and Ludvigson's model, from 57% to 9% for Lustig and Van Nieuwerburgh's model, from 41% to 3% for Santos and Veronesi's model, from 80% to 42% for Li, Vassalou, and Xing's model, and from 18% to 3% for Yogo's model. In addition, the χ^2 statistics are generally insignificant in tests with size-B/M portfolios but strongly reject using the expanded set of 55 portfolios.

The second key result is that the cross-sectional R^2 is often very uninformative about a model's true (population) performance. Our simulations show that, across the five models in Table 1, a 95% confidence interval for the true R^2 has an average width of 0.74 for the size-B/M portfolios and 0.73 for the expanded set of 55 portfolios. For regressions with size-B/M portfolios, we cannot reject that all models work perfectly, as expected, but neither can we reject that the true R^2s are quite small, with an average lower bound for the confidence intervals of 0.26. (Li, Vassalou, and Xing's model is an outlier, with a lower bound of 0.65.) For regressions with all 55 portfolios, the confidence intervals all include 0.00 or 0.05 – that is, using just the sample R^2, we can't reject that the models have essentially no explanatory power. Two of the confidence intervals cover almost the entire range of R^2s from 0.00 to 1.00. The table suggests, as a general rule, that sampling variation in the R^2 is just too large to use it as a reliable metric of performance.

[4] The only other difference is that, to simulate data for different true cross-sectional R^2s, we keep the true factor loadings the same in all simulations, equal to the historical estimates, and change the vector of true expected returns to give the right R^2. Specifically, expected returns in the simulations equal $\mu = h\,(C\,\lambda) + \varepsilon$, where C is the estimated matrix of factor loadings for a model, λ is the estimated vector of cross-sectional slopes, h is a scalar constant, and ε is a random drawn from a MVN[0, σ_ε^2 I] distribution; h and σ_ε are chosen to give the right cross-sectional R^2 and to maintain the historical cross-sectional dispersion in expected returns.

Finally, in the spirit of taking seriously the cross-sectional parameters (Prescription 2), the table shows that none of the models explains the level of expected returns: the estimated intercepts are all significantly greater than zero for tests with either the size-B/M portfolios or the expanded set of 55 portfolios. The regressions use excess quarterly returns (in %), so the intercepts can be interpreted as the estimated quarterly zero-beta rates over and above the riskfree rate. Annualized, the zero-beta rates range from 7.8% to 14.3% *above the riskfree rate*. These estimates cannot reasonably be attributed to differences in lending vs. borrowing costs.

In sum, despite the seemingly impressive ability of the models to explain the cross section of average returns on size-B/M portfolios, none of the models performs very well once we expand the set of test portfolios, consider sampling error in the R^2s, or ask them to price the riskfree asset.

5. Conclusion

The main point of the paper is easily summarized: Asset-pricing models should not be judged by their success in explaining cross-sectional variation in average returns for size-B/M portfolios (more generally, for portfolios in which a couple of factors are known to explain most of the time series and cross-sectional variation in returns). High cross-sectional explanatory power for size-B/M portfolios, in terms of high R^2 or small pricing errors, is simply not a sufficiently high hurdle. Further, the sample cross-sectional R^2 seems to be relatively uninformative about the true (population) performance of a model, at least in our tests with size, B/M, and industry portfolios.

The problems we highlight are not just sampling issues, i.e., they are not solved by getting standard errors right (but sampling issues do make them worse). In population, if returns have a factor structure like that of size-B/M portfolios, true expected returns will line up with true factor loadings so long as a proposed factor is correlated with returns only through the variation captured by the two or three common components in returns. The problems are also not solved by using an SDF-GMM approach, since SDF tests are very similar to traditional cross-sectional regressions.

The paper offers a number of suggestions for improving empirical tests. First, since the problems are tied to the strong covariance structure of size–B/M portfolios, one simple suggestion is to expand the set of test assets to include portfolios sorted in other ways, for example, by industry or factor loadings. Second, since the problems are exacerbated by the fact that empirical tests often ignore theoretical restrictions on the cross-sectional slopes, another suggestions is to take these restrictions seriously when theory provides appropriate guidance. Last, since the problems are exacerbated by sampling issues, a third suggestion is to report confidence intervals for the cross-sectional R^2 (one technique for doing so is described in the paper). Together, these prescriptions should help to improve the power and rigor of empirical tests, though they clearly don't provide a perfect solution.

Appendix

This appendix derives the small-sample distribution of the HJ-distance when returns are multivariate normal and the factors in the proposed model P are portfolio returns (or have been replaced by maximally correlated mimicking portfolios). R is defined, for the purposes of this appendix, to be the N+1 vector of total rates of return on the test assets, including the riskless asset.

Let $y = a + b\ P$. The HJ-distance is defined as $D = \min_m E[(m - y)^2]$, where m represents any well-specified SDF, i.e., any variable for which $E[m(1+R)] = 1$. Hansen and Jagannathan (1997) show that, if y is linear in asset returns (or is the projection of a non-return y onto the space of asset returns), then the m* which solves the minimization problem is linear in the return on the tangency portfolio, i.e., $m^* = v_0 + v_1 R_\tau$ for some constants v_0 and v_1, and $D = E[(m^* - y)^2]$.

The constants a and b are generally unknown and chosen to minimize D. Therefore, a and b solve $\min_{a,b} [E(m^* - y)^2] = \min_{a,b} [\mathrm{var}(m^* - a - b\ P) + E^2(m^* - a - b\ P)]$. The value of a doesn't affect the variance term, so the optimal choice of a must make $E[m^*] = E[a + b\ P]$, implying that $D = \min_b \mathrm{var}(m^* - a - b\ P)$. This observation means that D is the residual variance when m* is regressed on a constant and P or, equivalently, that D is v_1^2 times the residual variance when R_τ is regressed on a constant and P: $D = v_1^2\ \mathrm{var}(\omega)$, where ω is from the regression $R_\tau = a' + b'\ P + \omega$. The correlation between any portfolio and the tangency portfolio equals the ratio of their Sharpe measures, $\mathrm{cor}(R_x, R_\tau) = \theta_x/\theta_\tau$. Thus, to minimize var(ω), b′ must give the linear combination of P that has the maximum squared Sharpe ratio, denoted θ_P^2, from which it follows that $\mathrm{var}(\omega) = (1 - \theta_P^2 / \theta_\tau^2)\ \sigma_\tau^2$. Cochrane (2001) shows that $v_1 = -\mu_\tau / [\sigma_\tau^2 (1+r_f)]$, implying that $D = v_1^2\ \mathrm{var}(\omega) = (\theta_\tau^2 - \theta_P^2) / (1 + r_f)^2$. It follows that $D = \theta_z^2 / (1 + r_f)^2$, where $\theta_z^2 = \theta_\tau^2 - \theta_P^2$ can be thought of as the proposed model's unexplained Sharpe ratio.

The analysis above is cast in terms of population parameters, but equivalent results go through in sample, re-defining all quantities as sample moments. Thus, the estimated HJ-distance, d, is the difference between the sample squared Sharpe ratios of the ex post tangency portfolio and the portfolios in P, equal to the central quadratic in the GRS F-statistic, $\hat{a}'\ \Sigma_{OLS}^{-1}\ \hat{a}$. Following the discussion in Section 4, the sample HJ-distance is proportional to the GRS F-statistic, $d = c\ F\ [N\ (T - K - 1) / (T - N - K)]$, where $c = (1 + s_P^2) / T$ and s_P^2 is the sample counterpart to θ_P^2 (i.e., the sample maximum squared Sharpe ratio attainable from combinations of P). It follows immediately that, up to a constant of proportionally, d is non-central F with non-centrality parameter $\lambda = c^{-1}\ a'\ \Sigma^{-1}\ a = c^{-1}\ (1 + r_f)^2\ D$.

References

Acharya, Viral and Lasse Pedersen, 2005. Asset pricing with liquidity risk. *Journal of Financial Economics* 77 (2), 375-410.

Black, Fisher, 1972. Capital market equilibrium with restricted borrowing. *Journal of Business* 45, 444-454.

Black, Fisher, Michael C. Jensen, and Myron Scholes, 1972. The capital asset pricing model: Some empirical findings, in: *Studies in the theory of capital markets*, ed. by Michael C. Jensen. New York: Praeger.

Brennan, Michael, Ashley Wang, and Yihong Xia, 2004. Estimation and test of a simple model of intertemporal asset pricing. *Journal of Finance* 59, 1743-1775.

Campbell, John and Tuomo Vuolteenaho, 2004. Bad beta, good beta. *American Economic Review* 94, 1249-1275.

Cochrane, John, 1996. A cross-sectional test of an investment-based asset pricing model, *Journal of Political Economy* 104, 572-621.

Cochrane, John, 2001. *Asset Pricing*. Princeton University Press, Princeton, NJ.

Daniel, Kent and Sheridan Titman, 2005. Testing factor-model explanations of market anomalies. Working paper (Northwestern University, Evanston, IL).

Fama, Eugene and Kenneth French, 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33, 3-56.

Gibbons, Michael, Stephen Ross, and Jay Shanken, 1989. A test of the efficiency of a given portfolio. *Econometrica* 57, 1121-1152.

Hansen, Lars, 1982. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica* 50, 1029-1054.

Hansen, Lars, John Heaton, and Nan Li, 2005. Consumption strikes back? Measuring long-run risk. Working paper (University of Chicago, Chicago, IL).

Hansen, Lars and Ravi Jagannathan, 1997. Assessing Specification Errors in Stochastic Discount Factor Models. *Journal of Finance* 52, 557-590.

Heaton, John and Deborah Lucas, 2000. Asset pricing and portfolio choice: The role of entrepreneurial risk. *Journal of Finance* 55 (3), 1163-1198.

Jagannathan, Ravi and Zhenyu Wang, 1996. The conditional CAPM and the cross-section of stock returns. *Journal of Finance* 51, 3-53.

Kan, Raymond and Chou Zhang, 1999. Two-pass tests of asset pricing models with useless factors. *Journal of Finance* 54, 203-235.

Kandel, Shmuel and Robert Stambaugh, 1995. Portfolio inefficiency and the cross-section of expected returns. *Journal of Finance* 50, 157-184.

Kimmel, Robert, 2003. Risk premia in linear factor models: Theoretical and econometric issues. Working paper (Princeton University, Princeton, NJ).

Kullman, Cornelia, 2003. Real estate and its role in asset pricing. Working paper (University of British Columbia, Vancouver, BC).

Lettau, Martin and Sydney Ludvigson, 2001. Resurrecting the (C)CAPM: A cross-sectional test when risk premia are time-varying. *Journal of Political Economy* 109, 1238-1287.

Lewellen, Jonathan and Stefan Nagel, 2005. The conditional CAPM does not explain asset-pricing anomalies. *Journal of Financial Economics*, forthcoming.

Li, Qing, Maria Vassalou, and Yuhang Xing, 2005. Sector investment growth rates and the cross-section of equity returns. *Journal of Business*, forthcoming.

Lustig, Hanno and Stijn Van Nieuwerburgh, 2004. Housing collateral, consumption insurance, and risk premia. *Journal of Finance* 60 (3), 1167-1221.

Mehra, Rajnish and Edward Prescott, 1985. The equity premium: A puzzle. *Journal of Monetary Economics* 15, 145-161.

Nawalkha, Sanjay, 1997. A multibeta representation theorem for linear asset pricing theories, *Journal of Financial Economics* 46, 357-381.

Pastor, Lubos and Robert Stambaugh, 2003. Liquidity risk and expected stock returns. *Journal of Political Economy* 111, 642-685.

Parker, Jonathan and Christian Julliard, 2005. Consumption risk and the cross section of expected returns. *Journal of Political Economy* 113 (1), 185-222.

Petkova, Ralitsa, 2005. Do the Fama-French factors proxy for innovations in predictive variables? *Journal of Finance*, forthcoming.

Piazzesi, Monika, Martin Schneider, and Selale Tuzel, 2004. Housing, consumption, and asset pricing. Working paper (University of Chicago).

Reisman, Haim, 1992. Reference variables, factor structure, and the approximate multibeta representation. *Journal of Finance* 47, 1303-1314.

Roll, Richard and Stephen Ross, 1994. On the cross-sectional relation between expected returns and betas. *Journal of Finance* 49, 101-121.

Santos, Tano and Pietro Veronesi, 2005. Labor income and predictable stock returns. *Review of Financial Studies*, forthcoming.

Shanken, Jay, 1987. Multivariate proxies and asset pricing relations: Living with the Roll critique. *Journal of Financial Economics* 18, 91-110.

Shanken, Jay, 1990. Intertemporal asset pricing: An empirical investigation. *Journal of Econometrics* 45, 99-120.

Shanken, Jay, 1992a. On the current state of the arbitrage pricing theory, *Journal of Finance* 47, 1569-1574.

Shanken, Jay, 1992b. On the estimation of beta-pricing models. *Review of Financial Studies* 5, 1-34.

Shanken, Jay and Mark Weinstein, 2005. Economic forces and the stock market revisited. Working paper (Emory University, Atlanta, GA).

Stock, James, 1991. Confidence intervals for the largest autoregressive root in U.S. macroeconomic time series. *Journal of Monetary Economics* 28, 435-459.

Vassalou, Maria, 2003. News related to future GDP growth as a risk factor in equity returns. *Journal of Financial Economics* 68, 47-73.

Yogo, Motohiro, 2005. A consumption-based explanation of expected stock returns. *Journal of Finance*, forthcoming.

Table 1. Empirical tests of asset-pricing models, 1963 – 2004.

This table reports slopes, t-statistics (in parentheses), and adj. R^2s from cross-sectional regressions of average excess returns on estimated factor loadings for five models proposed in the literature. Returns are quarterly (%). The test assets are Fama and French's 25 size-B/M portfolios used alone (FF25) or together with Fama and French's 30 industry portfolios (FF25+30 ind). t-statistics are adjusted for estimation error in factor loadings (Shanken, 1992b). A 95% confidence interval for the true R^2 is reported in brackets, based on the approach described in the text. The χ^2 statistic, with p-value in brackets, tests whether residuals in the cross-sectional regression are all zero. The models are estimated from 1963 – 2004 except Yogo's, for which we have factor data until 2001.

Model	Test assets	Variables				Adj. R^2	χ^2
Lettau & Ludvigson (2001)		**const.**	**cay**	**Δc**	**cay$\times\Delta$c**		
Conditional CCAPM	FF25	3.33	-0.81	0.25	0.00	0.58	34.9
		(2.31)	(1.26)	(0.85)	(0.43)	[0.30, 1.00]	[p=0.03]
	FF25+30 Ind	2.50	-0.48	0.09	0.00	0.00	198.8
		(3.32)	(1.24)	(0.53)	(1.11)	[0.00, 0.55]	[p=0.00]
Lustig & Van Nieuwerburgh (2004)		**const.**	**my**	**Δc**	**my$\times\Delta$c**		
Conditional CCAPM	FF25	3.58	4.23	0.02	0.10	0.57	21.5
		(2.25)	(0.78)	(0.04)	(1.59)	[0.25, 1.00]	[p=0.43]
	FF25+30 Ind	2.78	0.37	-0.02	0.03	0.09	161.5
		(3.54)	(0.13)	(0.09)	(1.41)	[0.00, 1.00]	[p=0.00]
Santos & Veronesi (2005)		**const.**	**R_M**	**$s^{\omega}\times R_M$**			
Conditional CAPM	FF25	2.46	-0.33	-0.22		0.41	20.1
		(1.41)	(0.18)	(2.00)		[0.10, 1.00]	[p=0.57]
	FF25+30 Ind	2.30	-0.17	-0.05		0.03	192.4
		(2.78)	(0.16)	(1.53)		[0.00, 0.60]	[p=0.00]
Li, Vassalou, & Xing (2005)		**const.**	**ΔI_{HH}**	**ΔI_{Corp}**	**ΔI_{Ncorp}**		
Investment growth model	FF25	2.47	-0.80	-2.65	-8.59	0.80	26.0
		(2.15)	(0.39)	(1.04)	(1.99)	[0.65, 1.00]	[p=0.21]
	FF25+30 Ind	2.22	0.20	-0.93	-5.10	0.42	145.3
		(3.18)	(0.19)	(0.59)	(2.34)	[0.05, 1.00]	[p=0.00]
Yogo (2005)		**const.**	**Δc_{Ndur}**	**Δc_{Dur}**	**R_M**		
Durables CCAPM	FF25	1.98	0.28	0.67	0.48	0.18	25.7
		(1.37)	(1.01)	(2.36)	(0.29)	[0.00, 1.00]	[p=0.21]
	FF25+30 Ind	1.95	0.18	0.19	0.12	0.03	163.9
		(2.30)	(1.02)	(1.53)	(0.11)	[0.00, 0.55]	[p=0.00]

Variables:
R_M = CRSP value-weighted excess return
Δc = log consumption growth
cay = Lettau and Ludvigson's (2001) consumption-to-wealth ratio
my = Lustig and Van Nieuwerburgh's (2004) housing collateral ratio (based on mortgage data)
s^{ω} = labor income to consumption ratio
ΔI_{HH}, ΔI_{Corp}, ΔI_{Ncorp} = log investment growth for households, non-financial corporations, and the non-corporate sector
Δc_{Ndur}, Δc_{Dur} = Yogo's (2005) log consumption growth for non-durables and durables