

In-Sample vs. Out-of-Sample Tests of Stock Return Predictability in the Context of Data Mining

David E. Rapach
Albers School of Business and Economics
Seattle University
900 Broadway
Seattle, WA 98122-4340
Phone: 206-296-5705
Fax: 206-296-2486
E-mail: rapachd@seattleu.edu

Mark E. Wohar*
Department of Economics
University of Nebraska at Omaha
CBA 512K
Omaha, NE 68182-0286
Phone: 402-554-3712
Fax: 402-554-2853
E-mail: wohar@unomaha.edu

July 5, 2002 (Revised)

Abstract

In this paper, we undertake an extensive analysis of in-sample and out-of-sample tests of stock return predictability in an effort to better understand the nature of the empirical evidence on return predictability. We show that a number of financial variables appearing in the extant literature display both in-sample and out-of-sample predictive ability with respect to stock returns in annual data covering most of the twentieth century and postwar quarterly data. In contrast to the extant literature, we demonstrate that there is not a great deal of discrepancy between the results from in-sample and out-of-sample tests, once we employ recently developed out-of-sample tests with good power properties. According to the conventional wisdom, our out-of-sample results suggest that the in-sample evidence of stock return predictability is not spurious, as out-of-sample tests of predictability help guard against data mining. However, Inoue and Kilian (2002) recently challenge the conventional wisdom and argue that out-of-sample tests are just as susceptible to data mining biases as in-sample tests. With this in mind, we test for return predictability using a bootstrap procedure that explicitly accounts for data mining when calculating critical values. We still find that certain financial variables display significant in-sample and out-of-sample predictive ability with respect to stock returns. Given that we obtain evidence of predictability in an econometric environment that controls for data mining, our in-sample and out-of-sample results strengthen the case for a predictable component in stock returns.

JEL classifications: C22, C52, C53, G12, G14

Key words: Stock return predictability; Nested models; In-sample tests; Out-of-sample tests; Data mining

*Corresponding author. The authors thank Todd Clark, Greg Ellis, Lutz Kilian, Mike McCracken, and Chris Weber for very helpful comments on earlier drafts. The usual disclaimer applies. The results reported in this paper were generated using GAUSS 3.6.

1. Introduction

There now exists a voluminous literature on the predictability of stock and bond returns from past information. Interestingly, researchers have identified a large number of financial variables that appear to predict future stock returns. These include the dividend-price ratio (Rozeff, 1984; Campbell and Shiller, 1988a; Fama and French, 1988; Hodrick, 1992), price-earnings ratio (Campbell and Shiller, 1988b, 1998), book-to-market ratio (Kothari and Shanken, 1997; Pontiff and Schall, 1998), market value-to-net worth ratio or “Fed q” (Robertson and Wright, 2000; Smithers and Wright, 2000), dividend-payout ratio (Lamont, 1998), term and default spreads on bonds (Campbell, 1987; Fama and French, 1989), short-term interest rate (Campbell, 1987; Hodrick, 1992; Ang and Bekaert, 2001), equity share in total new equity and debt issues (Baker and Wurgler, 2000), and consumption-wealth ratio (Lettau and Ludvigson, 2001).¹ The evidence for the predictability of stock returns comes primarily from in-sample predictive regression models. While there are important econometric difficulties relating to the lack of exogenous regressors and overlapping observations in predictive regression models (Mankiw and Shapiro, 1986; Stambaugh, 1986, 1999; Richardson and Stock, 1989; Nelson and Kim, 1993), Campbell (2000, p. 1523) nevertheless concludes, “Despite these difficulties, the evidence for predictability survives at reasonable if not overwhelming levels of statistical significance. Most financial economists appear to have accepted that aggregate returns do contain an important predictable component.”

While Campbell (2000) is probably correct in assessing the extant literature, as noted above, the extant literature primarily relies on in-sample tests of stock return predictability. This raises concerns of data mining, also referred to as model overfitting or data snooping. Lo and MacKinlay (1990) and Foster, Smith, and Whaley (1997) provide theoretical analyses of data mining in the context of return predictability. It is typically believed that out-of-sample tests provide a measure of protection against data mining, as statistical models are tested using out-of-sample observations that are not used in the estimation of the statistical model itself. It is interesting to note that the relatively few studies that employ

¹ The studies cited are representative and do not constitute an exhaustive list.

out-of-sample tests of return predictability typically obtain negative results. For example, in an effort to guard against model overfitting, Bossaerts and Hillion (1999) use different model selection criteria to choose the best forecasting model for real stock returns for a number of industrialized countries over the postwar period. Testing for out-of-sample forecasting power by regressing actual returns on the forecasts from the best models, they find that the best forecasting models for the United States fail to have significant out-of-sample forecasting power for S&P 500 excess returns at the 1-month horizon over the 1990:06-1995:05 out-of-sample period. They thus conclude that there is no external validation of the best forecasting models. Goyal and Welch (2002) also employ out-of-sample tests. They examine the predictive ability of the dividend-price ratio for CRSP value-weighted annual real returns in excess of the annual real returns on a 3-month Treasury bill over the 1926-2000 period. While they find evidence of in-sample predictability, the dividend-price ratio exhibits little out-of-sample predictive ability compared to a model of constant returns according to the Diebold and Mariano (1995) and West (1996) statistic.² The negative results typically generated by out-of-sample tests suggest that the in-sample evidence of return predictability is spurious.

The disparities between in-sample and out-of-sample test results of return predictability in the extant literature make an overall assessment of return predictability difficult. In this paper, we undertake an extensive analysis of both in-sample and out-of-sample tests of stock return predictability in an effort to better understand the empirical evidence on return predictability. We test whether the ten financial variables cited in the opening paragraph exhibit significant in-sample and out-of-sample predictive ability with respect to stock returns. We follow the extant literature and assess in-sample predictability via the t -statistic corresponding to the slope coefficient in a predictive regression model. In order to compare out-of-sample forecasts from a model of constant returns to forecasts from a model that utilizes a financial variable, we use two recently developed test statistics. The first, due to McCracken (2000), is a variant of the Diebold and Mariano (1995) and West (1996) statistic that tests for equal predictive ability. We also use a statistic

² Goyal and Welch (2002) use the standard normal as the limiting distribution for the Diebold and Mariano (1995) and West (1996) statistic. However, Clark and McCracken (2001a) show that this statistic has a non-standard limiting distribution when comparing forecasts from nested models, as Goyal and Welch (2002) do.

designed to test for forecast encompassing, a variant of the Harvey, Leybourne, and Newbold (1998) statistic due to Clark and McCracken (2001a). These variants have two key advantages over the original statistics. First, unlike the original statistics, the variants take into account the parameter estimation uncertainty associated with the construction of forecasts. Second, Clark and McCracken (2001a,b) find the variants to be considerably more powerful than the original statistics in extensive Monte Carlo simulations. These more powerful tests may be better equipped to detect out-of-sample predictability than the statistics typically used in the extant literature. The McCracken (2000) and Clark and McCracken (2001a) statistics have standard limiting distributions when comparing forecasts from non-nested models. However, tests of stock return predictability typically compare forecasts from nested models, as the competing model in the extant literature is usually one of constant expected returns. Clark and McCracken (2001a,b) show that the above statistics have non-standard limiting distributions when comparing forecasts from nested models. They also show that the limiting distribution is non-pivotal at horizons beyond one-period-ahead when comparing forecasts from nested models. It can thus be highly misleading to base inferences on standard asymptotic distributions when comparing forecasts from nested models. We follow the recommendation of Clark and McCracken (2001b) and base our inferences on a bootstrap procedure similar to the one used in Kilian (1999).³

Using annual data from 1927-1999, we find that the equity share has significant in-sample predictive ability at the 1-year horizon and that the term spread has significant in-sample predictive ability at the 5-year horizon with respect to S&P 500 real stock returns. In addition, the price-earnings ratio and Fed q have significant in-sample predictive ability at the 10-year horizon. At least one of the McCracken (2000) and Clark and McCracken (2001a) statistics also provide significant evidence of out-of-sample predictive ability for these same variables over the 1964-1999 out-of-sample period. For this data set, there is no discrepancy between the in-sample and out-of-sample test results once we use the most powerful out-of-sample tests. When we measure real returns based on the CRSP equal-weighted index, we identify four

³ We also base our in-sample inferences on the bootstrap procedure in order to control for the potential biases due to regressor endogeneity and overlapping observations noted in the opening paragraph.

variables—book-to-market ratio, Fed q , default spread, and equity share—with significant in-sample predictive ability at the 1-year horizon, while the dividend-price ratio and equity share display significant in-sample predictive ability at the 5-year horizon. In addition, the short-term interest rate shows significant in-sample predictive ability at the 10-year horizon. Some of these variables also display significant out-of-sample predictive ability. Overall, we find relatively little discrepancy between the results from in-sample and out-of-sample tests of predictability for our annual data.

We also test for stock return predictability using postwar quarterly data from 1953:2-2000:4. For S&P 500 returns, the consumption-wealth ratio demonstrates significant in-sample predictive ability at horizons of 1, 8, and 16 quarters; the term spread at horizons of 1 and 16 quarters; the default spread at the 1-quarter horizon; and the short-term interest rate at horizons of 1 and 8 quarters. For the consumption-wealth ratio and short-term interest rate, there is also significant evidence of out-of-sample predictive ability over the 1990:2-2000:4 period (selected to match the beginning of the out-of-sample period in Bossaerts and Hillion, 1999). Using CRSP equal-weighted returns, the dividend-price ratio, consumption-wealth ratio, term spread, default spread, and short-term rate all show significant in-sample predictive ability at the 1-quarter horizon, while the consumption-wealth ratio also displays significant in-sample predictive ability at horizons of 8 and 16 quarters. The consumption-wealth ratio, term spread, and short-term interest rate also exhibit significant out-of-sample predictive ability over the 1990:2-2000:4 period. Like the results for our annual data, we do not find large discrepancies between the results from in-sample and out-of-sample tests of predictability for our postwar quarterly data, once we use relatively powerful out-of-sample tests.

According to “conventional wisdom,” our out-of-sample evidence of stock returns predictability in annual and quarterly data is more reliable than our in-sample evidence, as it is less susceptible to data mining. However, Inoue and Kilian (2002) challenge the conventional wisdom and argue that out-of-sample tests should not be preferred to in-sample tests, as out-of-sample tests are no less susceptible to data mining than in-sample tests. Indeed, Inoue and Kilian (2002) show that, *if appropriate critical value are used*, in-sample and out-of-sample tests of predictability are equally reliable against data mining under the null

hypothesis of no predictability.⁴ Given the findings in Inoue and Kilian (2002), we compute appropriate critical values for all of our in-sample and out-of-sample statistics based on a data-mining bootstrap procedure. Even when we explicitly take data mining into account through the data-mining bootstrap, we still identify some financial variables with significant in-sample and out-of-sample predictive ability in both our annual and quarterly data. Overall, our results strengthen the case for the stock return predictability.

As a cautionary note, we emphasize that we are concerned with testing for the existence of return predictability in population. This is a conceptually and practically distinct issue from whether a practitioner in real time could have constructed a portfolio that earns extra-normal returns. A practitioner is interested in ranking models—without necessarily caring about the significance of any differences—and selecting the best forecasting model according to a profit-based metric.⁵ In contrast, predictability tests are designed to test for the existence of a predictive relationship in population.⁶

The rest of the paper is organized as follows. Section 2 presents our econometric methodology; Section 3 reports the in-sample and out-of-sample predictability test results for our annual and quarterly data; Section 4 summarizes our main findings.

2. Econometric Methodology

2.1. In-Sample Predictability

Extant studies of the predictability of stock returns are typically predicated on the following predictive regression model:

$$y_{t+k} = \alpha + \beta \cdot x_t + u_{t+k} , \quad (1)$$

⁴ Inoue and Kilian (2002) actually favor in-sample over out-of-sample tests of predictability, as in-sample tests typically are more powerful when both kinds of tests use appropriate critical values.

⁵ See Pesaran and Timmerman (1995) and Cooper, Gutierrez, and Marcum (2001) for examples of studies that examine whether portfolios that earn extra-normal returns could have been constructed in real time. In addition, we note that, as emphasized by Fama (1991), return predictability in population does not imply that markets are inefficient, as time-varying returns may be an equilibrium phenomenon.

⁶ See Lettau and Ludvigson (2002b) and Inoue and Kilian (2002) for more discussion on testing for predictability in population.

where y_{t+k} is the log real return to holding stocks from period t to $t+k$, x_t is a financial variable that is believed to potentially predict future real returns, and u_{t+k} is a disturbance term. The log real return to stocks, y_{t+k} , is typically measured by $y_{t+k} = \sum_{i=1}^k r_{t+i}$, where $r_t = \log[(S_t + D_t)/S_{t-1}] - \log(P_t/P_{t-1})$, S_t is the nominal stock price at the end of period t (or beginning of period $t+1$), D_t is dividends paid out during period t , and P_t is the aggregate price level at the end of period t (or beginning of period $t+1$). Suppose that we have observations for r_t and x_t for $t=1, \dots, T$. This leaves us with $T-k$ usable observations with which to estimate the in-sample predictive regression model. The predictive ability of x_t is typically assessed by examining the t -statistic corresponding to $\hat{\beta}$, the OLS estimate of β in equation (1), as well as the goodness-of-fit measure, R^2 . Under the null hypothesis of no predictability, $\beta=0$, so that expected returns are constant.⁷ Theory typically suggests the direction of the effect of x_t on y_{t+k} under the alternative hypothesis. We specify all of the financial variables in such a way that $\beta > 0$ under the alternative hypothesis. Inoue and Kilian (2002) recommend using theory to specify a one-sided alternative hypothesis for in-sample tests, as this incorporates economic content that makes for more powerful tests.

A well-known potential problem with estimating a predictive regression model such as equation (1) is small-sample bias, as x_t is not an exogenous regressor in equation (1). Stambaugh (1986, 1999) considers the case where $k=1$. He examines a bivariate system where the data-generating process (DGP) for y_{t+1} is governed by equation (1), and the DGP for x_t is a stationary first-order autoregressive (AR) process:

$$y_{t+1} = \alpha + \beta \cdot x_t + u_{t+1}, \quad (2)$$

$$x_{t+1} = \theta + \rho \cdot x_t + v_{t+1}, \quad (3)$$

⁷ Following Baker and Wurgler (2000), we first divide x_t by its standard deviation over the full sample. This normalization has no effect on in-sample or out-of-sample statistical inferences, but it makes it easier to compare the estimated β coefficient in equation (1) across financial variables, as the coefficient can be interpreted as the change in expected returns given a one-standard-deviation change in the financial variable.

with $-1 < \rho < 1$ ($0 < \rho < 1$ is typically the relevant range for financial variables). He shows that the bias in the OLS estimate of β is given by:

$$E(\hat{\beta} - \beta) \approx (\sigma_{uv} / \sigma_v^2) E(\hat{\rho} - \rho), \quad (4)$$

where $\hat{\rho}$ is the OLS estimate of ρ in equation (3), σ_{uv} is the covariance between u_{t+1} and v_{t+1} , and σ_v^2 is the variance of v_{t+1} . It is well-known that $\hat{\rho}$ is a biased estimator of ρ and that the bias is given by (Kendall, 1954; Shaman and Stine, 1988):

$$E(\hat{\rho} - \rho) \approx -(1 + 3\rho)/(T - 1). \quad (5)$$

Combining equations (4) and (5), we have:

$$E(\hat{\beta} - \beta) \approx -(\sigma_{uv} / \sigma_v^2) [(1 + 3\rho)/(T - 1)]. \quad (6)$$

From equation (6), we see that a persistent x_t series (a “large” ρ) and strong correlation between the residuals across equations (2) and (3) can result in a substantial finite-sample bias in $\hat{\beta}$. We are especially likely to encounter a substantial bias if x_t is a valuation ratio, such as the dividend-price ratio, as valuation ratios tend to be quite persistent and strongly correlated with returns. The use of valuation ratios to predict real returns is popular in the extant literature.

Additional potential problems emerge when $k > 1$ in equation (1). As is well-known, the observations for the regressand in equation (1) are overlapping when $k > 1$ and thus not independent (Richardson and Stock, 1989). This induces serial correlation in the disturbance term, u_{t+k} , and the standard errors used in constructing t -statistics need to account for this. A common procedure in the extant literature is to use Newey and West (1987) standard errors, as these are robust to heteroskedasticity and serial correlation in the disturbance term. (In our applications in Section 3 below, we use the Bartlett kernel and a lag truncation parameter of $[1.5 \cdot k]$, where $[\bullet]$ is the nearest integer function, when calculating Newey and West, 1987 standard errors.) However, even when robust standard errors are used to compute t -statistics, there is a strong tendency for the t -statistic corresponding to $\hat{\beta}$ to increase with the horizon under

the null hypothesis that $\beta = 0$ (Nelson and Kim, 1993; Goetzmann and Jorion, 1993; Kirby, 1997). Overall, the bias pointed out by Stambaugh (1986, 1999) and the overlapping nature of the observations at horizons beyond one period have the effect of shifting the distribution of the t -statistic for $\hat{\beta}$ to the right, sometimes severely, even when the t -statistics are based on Newey and West (1987) standard errors. Basing inferences on standard asymptotic distributions can thus lead to considerable size distortions when testing the null hypothesis of no predictability ($\beta = 0$). Given these potential problems, we follow much of the recent literature and conduct inference using a bootstrap procedure similar to the procedures in Nelson and Kim (1993), Mark (1995), Kothari and Shanken (1997), and Kilian (1999). The bootstrap procedure is described in detail below in Section 2.3.

2.2. Out-of-Sample Predictability

We next describe the recursive scheme used to generate out-of-sample predictions for y_{t+k} . First, we divide the total sample of T observations into in-sample and out-of-sample portions, where the in-sample observations span the first R observations for r_t and x_t . The first out-of-sample forecast for the “unrestricted” predictive regression model, equation (1), is generated in the following manner. Estimate the unrestricted predictive regression model via OLS using data available through period R ($x_1, \dots, x_{R-k}; y_{1+k}, \dots, y_{(R-k)+k}$). Armed with the OLS parameter estimates from equation (1) and x_R , construct a forecast for y_{R+k} based on the unrestricted predictive regression model using $\hat{y}_{1,R+k} = \hat{\alpha}_{1,R} + \hat{\beta}_{1,R} \cdot x_R$, where $\hat{\alpha}_{1,R}$ and $\hat{\beta}_{1,R}$ are the OLS estimates of α and β in equation (1) using data available through period R . Denote the forecast error by $\hat{u}_{1,R+k} = y_{R+k} - \hat{y}_{1,R+k}$. The initial forecast for the “restricted” predictive model is generated in a similar manner, except that we set $\beta = 0$ in equation (1). That is, we estimate the restricted regression model, equation (1) with $\beta = 0$, via OLS using data available through period R in order to form the forecast, $\hat{y}_{0,R+k} = \hat{\alpha}_{0,R}$, where $\hat{\alpha}_{0,R}$ is the OLS estimate of α in

equation (1) with β restricted to zero using data available through period R . Denote the forecast error corresponding to the restricted model as $\hat{u}_{0,R+k} = y_{R+k} - \hat{y}_{0,R+k}$.

In order to generate a second set of forecasts, we update the above procedure one period by using data available through period $R+1$. That is, we estimate the unrestricted and restricted regression models using data available through period $R+1$, and we use these parameter estimates and the observation for x_{R+1} in order to form unrestricted and restricted model forecasts for $y_{(R+1)+k}$ and their respective forecast errors, $\hat{u}_{1,(R+1)+k}$ and $\hat{u}_{0,(R+1)+k}$. We repeat this process through the end of the available sample, leaving us with two sets of $T-R-k+1$ recursive forecast errors, one each for the unrestricted and restricted regression models ($\{\hat{u}_{1,t+k}\}_{t=R}^{t=T-k}$ and $\{\hat{u}_{0,t+k}\}_{t=R}^{t=T-k}$).

The next step is to compare the out-of-sample forecasts from the unrestricted and restricted regression models. If the unrestricted model forecasts are superior to the restricted model forecasts, then the financial variable, x_t , improves the out-of-sample forecasts of y_{t+k} relative to a model of constant expected returns. A simple metric for comparing forecasts is Theil's U , the ratio of the root-mean-squared errors for the unrestricted and restricted regression model forecasts. If the mean-squared error (MSE) for the unrestricted model forecasts is less than the MSE for the restricted model forecasts, then $U < 1$. In order to test whether the unrestricted regression model forecasts are significantly superior to the restricted model forecasts, we use the McCracken (2000) $MSE-F$ statistic, a variant of the Diebold and Mariano (1995) and West (1996) statistic designed to test for equal predictive ability. We also use the Clark and McCracken (2001a) $ENC-NEW$ statistic, a variant of the Harvey, Leybourne, and Newbold (1998) statistic designed to test for forecast encompassing.

The $MSE-F$ statistic is used to test the null hypothesis that the unrestricted model forecast MSE is equal to the restricted model forecast MSE against the one-sided (upper-tail) alternative hypothesis that the unrestricted model forecast MSE is less than the restricted model forecast MSE. Like the Diebold and Mariano (1995) and West (1996) statistic, the $MSE-F$ statistic is based on the loss differential,

$$\hat{d}_{t+k} = (\hat{u}_{0,t+k})^2 - (\hat{u}_{1,t+k})^2. \quad \text{Let} \quad \bar{d} = (T - R - k + 1)^{-1} \sum_{t=R}^{T-k} \hat{d}_{t+k} = MSE_0 - MSE_1, \quad \text{where}$$

$MSE_i = \sum_{t=R}^{T-k} (\hat{u}_{i,t+k})^2$, $i = 0, 1$. The McCracken (2000) *MSE-F* statistic is given by:

$$MSE - F = (T - R - k + 1) \cdot \bar{d} / MSE_1. \quad (7)$$

A significant *MSE-F* statistic indicates that the unrestricted model forecasts are statistically superior to those of the restricted model. When comparing forecasts from nested models (as we do) and for $k=1$, McCracken (2000) shows that the *MSE-F* statistic has a non-standard limiting distribution that is pivotal and a function of stochastic integrals of Brownian motion.⁸ Clark and McCracken (2001b) demonstrate that the *MSE-F* statistic has a non-standard and non-pivotal limiting distribution in the case of nested models and $k > 1$. Given this last result, Clark and McCracken (2001b) recommend basing inference on a bootstrap procedure along the lines of Kilian (1999). Following this recommendation, we base our inferences on the bootstrap procedure described below in Section 2.3.

Our other statistic, *ENC-NEW*, relates to the concept of forecast encompassing.⁹ Forecast encompassing is based on optimally constructed composite forecasts. If the forecasts from the restricted regression model encompass the unrestricted model forecasts, the financial variable provides no useful additional information for predicting returns relative to a model of constant expected returns. If we reject forecast encompassing, then the financial variable does contain information useful for predicting returns apart from a model of constant returns. The Clark and McCracken (2001a) *ENC-NEW* statistic is a variant of the Harvey, Leybourne, and Newbold (1998) statistic, and it takes the form:

$$ENC - NEW = (T - R - k + 1) \cdot \bar{c} / MSE_1, \quad (8)$$

⁸ While West (1996) shows that the Diebold and Mariano (1995) and West (1996) statistic has a standard normal limiting distribution when comparing forecasts from non-nested models, McCracken (2000) shows that it has a non-standard limiting distribution when comparing forecasts from nested models.

⁹ See Clements and Hendry (1998) for a textbook discussion of forecast encompassing.

where $\hat{c}_{t+k} = \hat{u}_{0,t+k}(\hat{u}_{0,t+k} - \hat{u}_{1,t+k})$ and $\bar{c} = (T - R - k + 1)^{-1} \sum_{t=R}^{T-k} \hat{c}_{t+k}$. Under the null hypothesis, the restricted model forecasts encompass the unrestricted model forecasts, while under the one-sided (upper-tail) alternative hypothesis, the restricted model forecasts do not encompass the unrestricted model forecasts. Similar to the *MSE-F* statistic, the limiting distribution of the *ENC-NEW* statistic is non-standard and pivotal for $k=1$ (Clark and McCracken, 2001a), while it is non-standard and non-pivotal for $k > 1$ (Clark and McCracken, 2001b) when comparing forecasts from nested models.¹⁰ Again, Clark and McCracken (2001b) recommend basing inference on a bootstrap procedure given the non-pivotal limiting distribution.

As noted in the introduction, the *MSE-F* and *ENC-NEW* statistics have two key advantages over the original Diebold and Mariano (1995) and West (1996) and Harvey, Leybourne, and Newbold (1998) statistics. First, they account for the parameter uncertainty inherent in estimating the unrestricted and restricted models that are used to form the competing forecasts. Second, Clark and McCracken (2001a,b) find that the *MSE-F* and *ENC-NEW* statistics have good size properties and are typically more powerful than the original statistics in extensive Monte Carlo simulations with nested models. The simulations in Clark and McCracken (2001a,b) are especially important for the present paper, as they are based on bivariate vector autoregressions that are similar in structure to the system defined by equations (2) and (3).

¹⁰ As pointed out by Clark and McCracken (2001b), the Harvey, Leybourne, and Newbold (1998) statistic has a standard normal limiting distribution when comparing forecasts from non-nested models according to the theory in West (1996). However, for nested models, Clark and McCracken (2001a,b) show that it has a non-standard limiting distribution

2.3. Bootstrap Procedure

For the reasons discussed in Sections 2.1 and 2.2, we base inferences on a bootstrap procedure similar to the procedures in Nelson and Kim (1993), Mark (1995), Kothari and Shanken (1997), and Kilian (1999). We postulate that the data are generated by the following system under the null hypothesis of no predictability:

$$r_t = a_0 + \varepsilon_{1,t}, \quad (9)$$

$$x_t = b_0 + b_1 \cdot x_{t-1} + \dots + b_p \cdot x_{t-p} + \varepsilon_{2,t}, \quad (10)$$

where the disturbance vector, $\varepsilon_t = (\varepsilon_{1,t}, \varepsilon_{2,t})'$, is independently and identically distributed with covariance matrix Σ . We first estimate equations (9) and (10) via OLS, with the lag order (p) in equation (10) selected using the AIC (considering a maximum lag order of four),¹¹ and compute the OLS residuals, $\{\hat{\varepsilon}_t = (\hat{\varepsilon}_{1,t}, \hat{\varepsilon}_{2,t})'\}_{t=1}^{T-p}$. In order to generate a series of disturbances for our pseudo-sample, we randomly draw (with replacement) $T+100$ times from the OLS residuals, $\{\hat{\varepsilon}_t\}_{t=1}^{T-p}$, giving us a pseudo-series of disturbance terms, $\{\hat{\varepsilon}_t^*\}_{t=1}^{T+100}$. Note that we draw from the OLS residuals in tandem, thus preserving the contemporaneous correlation between the disturbances in the original sample. Denote the OLS estimate of a_0 in equation (11) by \hat{a}_0 , the OLS estimate of b_0 in equation (10) by \hat{b}_0 , and the bias-adjusted OLS estimates of (b_1, \dots, b_p) by $(\hat{b}_1^A, \dots, \hat{b}_p^A)$, where we use the bias adjustments in Shaman and Stine (1988, Table 1). Using $\{\hat{\varepsilon}_t^*\}_{t=1}^{T+100}$, $(\hat{a}, \hat{b}_0, \hat{b}_1^A, \dots, \hat{b}_p^A)$ in equations (9) and (10), and setting the initial observations for $(x_{t-1}, \dots, x_{t-p})$ equal to zero in equation (10), we can build up a pseudo-sample of $T+100$ observations for r_t and x_t , $\{r_t^*, x_t^*\}_{t=1}^{T+100}$. We drop the first 100 transient start-up observations in order to randomize the initial $(x_{t-1}, \dots, x_{t-p})$ observations, leaving us with pseudo-sample of T observations, matching the original sample. For the pseudo-sample, we calculate the t -statistic corresponding to β in the in-sample predictive regression model, equation (1), and the two out-of-sample statistics outlined in Section 2.2. We repeat this

¹¹ Kilian (1999) recommends using the AIC to select the lag order in equation (10).

process 1,000 times, giving us an empirical distribution for the in-sample t -statistic and each of the out-of-sample statistics. For each of the statistics, the p -value is the proportion of the bootstrap statistics that are greater than the statistic computed using the original sample.

2.4. Data-Mining Bootstrap Procedure

In Section 3 below, we use the in-sample and out-of-sample statistics described above to test the ability of nine financial variables to predict real stock returns using a long span of annual data. We also test the ability of eight financial variables to predict real stock returns using postwar quarterly data. When testing the predictive ability of a large number of financial variables, data mining becomes a serious concern. Lo and MacKinlay (1990) and Foster, Smith, and Whaley (1997) point this out with respect to in-sample tests of security return predictability. While data mining is generally considered to be a serious problem for in-sample tests of predictability, Inoue and Kilian (2002, p. 4) observe that “[t]here is a common perception that in the presence of data mining out-of-sample tests of predictability are more reliable than in-sample tests.” However, they argue,

Contrary to the conventional wisdom, [out-of-sample tests] offer no more protection from data mining than in-sample tests of predictability. The problem is that...the researcher knows exactly the ‘out-of-sample’-performance of any given model and is free to experiment with alternative predictors...Thus, out-of-sample inference is subject to exactly the same potential data mining problems as in-sample inference (Inoue and Kilian, 2002, p. 4).

As emphasized by Inoue and Kilian (2002), the key to controlling for data mining is the use of appropriate critical values for both in-sample and out-of-sample predictability tests. They consider a data-mining environment that is relevant for our predictability tests in the present paper. Suppose we consider J different financial variables as candidate predictors in the predictive regression model, equation (1): $x_{j,t}, j = 1, \dots, J$. The bootstrap procedure above implicitly assumes that we analyze each financial variable in isolation. In actuality, however, we consider a large number of potential predictors, but focus on the

variables that give the “best” results. In order to account for data mining when testing predictability, Inoue and Kilian (2002) specify the null hypothesis as $H_0 : \beta_j = 0 \forall j$ and the alternative hypothesis as $H_1 : \beta_j > 0$ for some j , where β_j is the slope coefficient in equation (1) when the explanatory variable is $x_{j,t}$. For an in-sample test statistic, they suggest using $\max_{j \in \{1, \dots, J\}} t_{\hat{\beta}_j}$, where $t_{\hat{\beta}_j}$ is the t -statistic corresponding to $\hat{\beta}_j$. For out-of-sample tests statistics, they suggest using the maximal $MSE-F$ and maximal $MSE-NEW$ statistics. Inoue and Kilian (2002) derive the asymptotic distribution for the maximal in-sample and out-of-sample statistics under the null hypothesis of no predictability, as well as local alternatives, in this data-mining environment. The limiting distributions are generally data-dependent, making inferences based on asymptotic distributions difficult. Inoue and Kilian (2002) recommend that bootstrap procedures be used in practice.

We modify the bootstrap procedure in Section 2.3 above in order to explicitly account for data mining. Again, suppose we have J different financial variables serving as candidate predictors in the predictive regression model, equation (1): $x_{j,t}, j = 1, \dots, J$. We posit that the data are generated by the following system under the null hypothesis of no predictability:

$$r_t = a_0 + \varepsilon_{1,t}, \quad (11)$$

$$\begin{aligned} x_{1,t} &= b_{1,0} + b_{1,1} \cdot x_{1,t-1} + \dots + b_{1,p_1} \cdot x_{1,t-p_1} + \varepsilon_{1,2,t}, \\ &\vdots \\ x_{J,t} &= b_{J,0} + b_{J,1} \cdot x_{J,t-1} + \dots + b_{J,p_J} \cdot x_{J,t-p_J} + \varepsilon_{J,2,t}, \end{aligned} \quad (12)$$

where the disturbance vector, $\varepsilon_t = (\varepsilon_{1,t}, \varepsilon_{1,2,t}, \dots, \varepsilon_{J,2,t})'$, is independently and identically distributed with covariance matrix Σ . We first estimate equation (11) and each of the AR processes in equation (12) via OLS. Note that the lag order for each of the AR processes in equation (12) can differ, and we select each lag order using the AIC (considering a maximum lag order of four).¹² We then compute the OLS residuals, $\{\hat{\varepsilon}_t = (\hat{\varepsilon}_{1,t}, \hat{\varepsilon}_{1,2,t}, \dots, \hat{\varepsilon}_{J,2,t})'\}_{t=1}^{T-p}$, where $p = \max_{j \in \{1, \dots, J\}} p_j$. In order to generate a series of disturbances for

¹² In order to conserve degrees of freedom, we do not estimate a VAR process for $(x_{1,t}, \dots, x_{J,t})$.

our pseudo-sample, we randomly draw (with replacement) $T + 100$ times from the OLS residuals, $\{\hat{\varepsilon}_t\}_{t=1}^{T-p}$, giving us a pseudo-series of disturbance terms, $\{\hat{\varepsilon}_t^*\}_{t=1}^{T+100}$. Drawing the OLS residuals in tandem preserves the contemporaneous correlation between all of the disturbances in the original sample. Using the pseudo-series of disturbance terms, the OLS estimates of the coefficients in equations (11) and (12),¹³ and setting the initial observations for each of the $x_{j,t}$ processes equal to zero in equation (12), we can build up a pseudo-sample of $T + 100$ observations for r_t and $x_{1,t}, \dots, x_{J,t}$, $\{r_t^*, x_{1,t}^*, \dots, x_{J,t}^*\}_{t=1}^{T+100}$. We drop the first 100 transient start-up observations in order to randomize the initial observations, leaving us with pseudo-sample of T observations, matching the original sample. For the pseudo-sample, we calculate the t -statistic corresponding to β_j in the in-sample predictive regression model and the two out-of-sample statistics for each of the $x_{j,t}^*$ variables in turn. We then compute and store the maximal t -statistic, as well as the maximal *MSE-F* and *ENC-NEW* statistics. We repeat this process 1,000 times, giving us an empirical distribution for the maximal t -statistic and each of the maximal out-of-sample statistics. After ordering the empirical distribution for each maximal statistic, the 900th, 950th, and 990th values serve as the 10%, 5%, and 1% critical values for each maximal statistic.

3. Empirical Results

3.1. 20th Century Annual Data

We use annual observations from 1927-1999, based on data availability, for two stock return series, S&P 500 log real returns and CRSP equal-weighted log real returns, and nine financial variables. The nine financial variables all appear in the extant literature: dividend-price ratio, log-level; price-earnings ratio, log-level; book-to-market ratio, log-level; Fed q (market value-to-net worth ratio), log-level; payout (dividend-earnings) ratio, log-level; term spread; default spread; short-term interest rate; equity share. The financial variables are described in detail in the Data Appendix. The term spread is measured as the difference

¹³ We again use bias-adjusted slope coefficients for the AR processes in equation (12).

between long-term and short-term government bond yields. The default spread is the difference between the Moody's seasoned Baa corporate bond yield and the Moody's seasoned Aaa corporate bond yield. The equity share is the annual total of equity issues divided by the sum of the annual total of equity issues and the annual total of long-term debt issues. The Fed q (Robertson and Wright, 2000; Smithers and Wright, 2000) and equity share (Baker and Wurgler, 2000) have appeared only recently in the extant literature. Theory suggests that increases in the dividend-price ratio, book-to-market ratio, payout ratio, term spread, and default spread should increase future returns, while increases in the price-earnings ratio, Fed q, short-term interest rate, and equity share should decrease future returns. In order to express the null hypothesis as $\beta > 0$ in equation (1) for all of the financial variables, we take the negative of the latter four variables before they enter into equation (1). Descriptive statistics for the two returns and each financial variable are reported in Table 1. It is evident that the two return series are highly volatile and display only weak serial correlation. Apart from the term spread and equity share, the financial variables display considerable persistence.

Tables 2 and 3 report in-sample predictive regression estimation results for horizons of 1, 5, and 10 years for S&P 500 and CRSP equal-weighted log real returns. Given that we use annual data from 1927-1999, we have 72 usable observations when $k=1$ for the predictive regression model, equation (1). The total number of usable observations decreases by one for each unit increase in k , leaving 63 usable observations when $k=10$. The p -values for the in-sample t -statistics reported in Tables 2 and 3 are generated using the bootstrap procedure described in Section 2.3 above. From Table 2, we see that the only financial variable that is a significant in-sample predictor of real S&P 500 returns at the 1-year horizon is the equity share. At the 5-year horizon, the term spread is a significant predictor (at the 10% level) of real S&P 500 returns. At the long horizon of 10 years, there is more evidence of in-sample predictability. This is consistent with the pattern in the extant literature, where evidence of in-sample predictability increases with the horizon. Both the price-earnings ratio and Fed q are significant predictors at the 10-year horizon. Overall, we identify four variables that have significant in-sample

predictive ability for S&P 500 returns at some horizon in Table 2: price-earnings ratio, Fed q, term spread, and equity share.

Turning to the in-sample predictive regression estimation results for the CRSP equal-weighted log real returns in Table 3, four variables—book-to-market ratio, Fed q, default spread, and equity share—show significant in-sample predictive ability at the 1-year horizon at conventional significance levels. At the 5-year horizon, the dividend-price ratio and, again, the equity share have significant predictive power, while the term spread and short-term interest rate have significant predictive ability at the 10-year horizon. Interestingly, the pattern of predictive ability for the Fed q is quite different for the CRSP equal-weighted and S&P 500 returns. From Table 2, we see that the Fed q has significant in-sample predictive power at the long horizon of 10 years for S&P 500 returns, while from Table 3 we see that it has predictive power at the short horizon of 1 year for CRSP equal-weighted returns. Overall, we find significant evidence of in-sample predictive ability for a number of financial variables in Tables 2 and 3.

We next discuss the out-of-sample test results in Table 2 and 3. We test out-of-sample predictive ability using the *MSE-F* and *ENC-NEW* statistics described in Section 2.2 above. We first must decide on the sample-split parameter (R), and we face a tradeoff at this point. If we limit the out-of-sample forecasts to very recent periods, we have very few out-of-sample observations to use in calculating the out-of-sample test statistics. This makes our inferences regarding out-of-sample predictability less reliable. If we begin our out-of-sample forecasts very early in the sample, we do not have many in-sample observations available to estimate the predictive regression models used to generate the initial out-of-sample forecasts. As a reasonable compromise, we decide to reserve the first half of the sample to estimate the unrestricted and restricted predictive regression models used to form the initial recursive out-of-sample forecasts, so that our first out-of-sample forecasts are for 1964. This leaves us with 36 out-of sample observations to use in computing the out-of-sample test statistics for $k=1$ and 27 out-of-sample observations for $k=10$. Note that our out-of-sample forecasts cover widely varying market conditions, including the bull market of the 1960s, the bear market of the 1970s, the 1987 crash, and the recent bull market of the 1990s. As with the in-

sample tests, the p -values for the out-of-sample results reported in Tables 2 and 3 are generated using the bootstrap procedure described in Section 2.3 above.

The sense of the relatively few extant studies that consider out-of-sample tests, such as Bossaerts and Hillion (1999) and Goyal and Welch (2002), is that while a number of financial variables display significant in-sample predictive ability with respect to stock returns, there is very little evidence of return predictability using out-of-sample tests. With this in mind, the most striking result in Table 2 is that there is no disagreement between the in-sample and out-of-sample test results when we use the relatively powerful *MSE-F* and *ENC-NEW* out-of-sample statistics. Each instance of a significant in-sample t -statistic for annual S&P 500 returns is matched by significant out-of-sample *MSE-F* and *ENC-NEW* statistics. Turning to Table 3, we see that three of the eight instances of a significant in-sample t -statistic find at least some out-of-sample corroboration (Fed q and equity share at the 1-year horizon and the equity share at the 5-year horizon). It is interesting to observe that we have two instances where there is evidence of significant out-of-sample predictive ability even when there is no evidence of significant in-sample predictive ability (price-earnings ratio at the 1-year horizon and dividend-price ratio at the 10-year horizon).

Overall, the results in Tables 2 and 3 challenge the notion that while there may be considerable evidence of in-sample return predictability, this evidence is spurious due to a paucity of out-of-sample evidence. When we use the recently developed *MSE-F* and *ENC-NEW* out-of-sample tests of McCracken (2000) and Clark and McCracken (2001a), we do not find great disagreement between the results from in-sample and out-of-sample tests of return predictability.

While the results in Tables 2 and 3 indicate that a number of financial variables exhibit significant in-sample and out-of-sample predictive ability with respect to S&P 500 and/or CRSP equal-weighted returns, the consideration of a large number of financial variables renders the predictability tests susceptible to data mining. As stressed by Inoue and Kilian (2002), out-of-sample tests are no less susceptible to data mining than in-sample tests. To control for data mining, we use the data-mining environment and the bootstrap procedure described above in Section 2.4. In the data-mining environment, we test for stock return

predictability using the null hypothesis that $\beta = 0$ for all nine of the financial variables included in Tables 2 and 3 against the alternative hypothesis that $\beta > 0$ for at least one of the financial variables using the maximal in-sample t -statistic and the maximal out-of-sample $MSE-F$ and $ENC-NEW$ statistics. Critical values for the in-sample and two out-of-sample maximal statistics are reported in Table 4.

From the S&P 500 return results reported in Table 2, the maximal t -statistic at the 1-year horizon is 3.01, corresponding to the equity share. From Panel A in Table 4, we see that the in-sample maximal t -statistic is still significant at the 5% level when we use critical values that take data mining into account. The out-of-sample maximal $MSE-F$ statistic of 3.50 corresponding to the equity share is also significant at the 10% level according to the critical values in Table 4.¹⁴ In light of the critical values in Table 4, the predictive ability detected in Table 2 for the equity share does not appear attributable to data mining. At the 5-year horizon, the maximal t -statistic from Table 2 is 1.57, corresponding to the term spread. According to the critical values in Table 4, this maximal t -statistic is not significant when we account for data mining. In addition, neither of the maximal out-of-sample statistics from Table 2 are significant at the 5-year horizon using the critical values in Table 4. The evidence of S&P 500 return predictability at the 5-year horizon in Table 2 is not robust to an environment that accounts for data mining. At the 10-year horizon, the maximal t -statistic from Table 2 is 7.27, corresponding to the Fed q . This maximal t -statistic is not significant at conventional significance levels according to the data-mining robust critical values. Interestingly, however, the out-of-sample maximal $MSE-F$ statistic of 53.24 corresponding to the Fed q is significant at the 5% level using the data-mining robust critical values, and the maximal $ENC-NEW$ statistic of 44.29, again corresponding to the Fed q , is significant at the 10% level. Thus, when we control for data mining, we still find significant evidence of predictability at the 10-year horizon.

¹⁴ As noted in footnote 4, Inoue and Kilian (2002) show that in-sample tests of predictability are more powerful than out-of-sample tests when data-mining robust critical values are used. This is due, in part, to the fact that in-sample tests use economic theory to specify that $\beta_j > 0$ under the alternative hypothesis, while the out-of-sample tests do not place such a restriction on β_j under the alternative hypothesis. Given that in-sample tests are more powerful than out-of-sample tests when data-mining robust critical values are used, evidence of predictability according to out-of-sample tests can be construed as especially strong evidence of predictability under a broader interpretation of the conventional wisdom.

Data-mining robust critical values for the maximal statistics corresponding to CRSP equal-weighted returns are reported in Panel B of Table 4. From Table 3, the maximal t -statistic at the 1-year horizon is 3.89, again corresponding to the equity share. According to the critical values in Table 4, the maximal t -statistic is significant at the 1% level.¹⁵ (The maximal t -statistic corresponding to the Fed q , 2.81, is also significant at the 1-year horizon.) In addition, the maximal out-of-sample $MSE-F$ statistic of 4.03 corresponding to the equity share is significant at the 10% level, and the maximal $ENC-NEW$ statistic of 5.57 is significant at the 5% level. None of the in-sample or out-of-sample maximal statistics are significant at the 5-year and 10-year horizons, so that robust evidence of predictability for CRSP equal-weighted returns is limited to a short horizon.

3.2. Postwar Quarterly Data

In order to explore stock return predictability more extensively, we also perform in-sample and out-of-sample tests using postwar quarterly data covering 1953:2-2000:4. We again use S&P 500 and CRSP equal-weighted log real returns, and we consider eight financial variables: seven of the nine financial variables that we considered for our annual data, with the exceptions of the equity share and book-to-market ratio (for which quarterly data are not available), and the consumption-wealth ratio. For the quarterly data, the short-term interest rate is measured in deviations from a 1-year moving average, following the convention in the extant literature. The consumption-wealth ratio is from Lettau and Ludvigson (2001, 2002a). It is calculated as $cay = c - 0.2985 \cdot a - 0.597 \cdot y$, where c is real per-capita consumption of nondurables and services, a is

¹⁵ As noted in the introduction, Foster, Smith, and Whaley (1997) provide a theoretical analysis of data mining in predictive regression models. Based on the Bonferroni inequality, they develop a weaker bound for the distribution of the maximal R^2 statistic under the null hypothesis of no predictability when a total of m explanatory variables are considered in regression models that include q explanatory variables at a time; see equation (3) in Foster, Smith, and Whaley (1997, p. 595). In our applications in Tables 2 and 3, $m = 9$ and $q = 1$. According to this distribution, the maximal R^2 statistic of 0.11 at the 1-year horizon in Table 2 is significant at the 5% level (the 5% critical value is 0.105), and the maximal R^2 statistic of 0.18 at the 1-year horizon in Table 3 is significant at the 1% level (the 1% critical value is 0.140). There is thus significant evidence of predictability in Tables 2 and 3 after we control for data mining using the distribution in Foster, Smith, and Whaley (1997). As the distribution theory in Foster, Smith, and Whaley (1997) assumes that the disturbance term in equation (1) is serially uncorrelated, we cannot use it to make inferences for the maximal R^2 at the 5-year and 10-year horizons in Tables 2 and 3.

financial wealth, and y is labor income, all measured in natural logarithms. The coefficients are based on the estimated cointegrating vector in Lettau and Ludvigson (2002a). Descriptive statistics for the quarterly return series and financial variables are provided in Table 5. Similar to the annual data, we see that the quarterly return series are volatile and very weakly serially correlated, while the quarterly financial variables are less volatile and more persistent.

In-sample predictability test results for S&P 500 returns are reported in Table 6, and the results for CRSP equal-weighted returns are reported in Table 7. We consider horizons of 1, 8, and 16 quarters. For the predictive regression model, equation (1), we have 190 usable observations at the 1-quarter horizon and 175 at the 16-quarter horizon. From Table 6, we see that the t -statistic is significant at the 1-quarter horizon for the consumption-wealth ratio (cay), term spread, default spread, and short-term interest rate. At the 8-quarter horizon, the t -statistic is significant for cay and the short-term interest rate, while it is significant for cay and the term spread at the 16-quarter horizon. Turning to the in-sample results for CRSP equal-weighted returns in Table 7, the t -statistic is significant at the 1-quarter horizon for the dividend-price ratio, cay , term spread, default spread, and short-term interest rate. It is also significant at horizons of 8 and 16 quarters for cay .

For the out-of-sample tests with our quarterly data, we again have to select an out-of-sample period for forecasting. Our out-of-sample period begins in 1990:2, which corresponds to the beginning of the Bossaerts and Hillion (1999) out-of-sample period.¹⁶ With quarterly postwar data, we can focus on an out-of-sample period that begins in the relatively recent past and still have a reasonable number of out-of-sample forecasts on which to base the out-of-sample statistics. Given a 1990:2-2000:4 out-of-sample period, we have 43 out-of sample observations to use in computing the out-of-sample test statistics at the 1-quarter horizon and 28 out-of-sample observations at the 16-quarter horizon. From the results for S&P 500 returns in Table 6, we see that the *ENC-NEW* out-of-sample statistic corroborates the in-sample finding of predictive ability for cay at all three forecast horizons. The *MSE-F* statistic also corroborates the in-sample

¹⁶ Bossaerts and Hillion (1999) use monthly data and an out-of-sample period beginning in 1990:06.

finding for *cay* at the 8-quarter and 16-quarter horizons.¹⁷ However, while there is significant evidence of in-sample predictive ability for the term spread, default spread, and short-term interest rate in Table 6 at certain horizons, neither of the out-of-sample statistics is significant at any horizon for these three variables.

Turning to the results for CRSP equal-weighted returns in Table 7, the significant evidence of in-sample predictive ability at horizons of 1, 8, and 16 quarters for *cay* is accompanied by significant evidence of out-of-sample predictive ability at these three horizons according to the *ENC-NEW* statistic. Both the *MSE-F* and *ENC-NEW* statistics indicate significant out-of-sample predictive ability for the term spread at the 1-quarter horizon, matching the significant evidence of in-sample predictive ability for the term spread at the 1-quarter horizon. In addition, the significant evidence of in-sample predictive ability for the short-term interest rate at the 1-quarter horizon is matched by significant evidence of out-of-sample predictive ability according to the *ENC-NEW* statistic. For the dividend-price ratio and default spread, there is no significant evidence of out-of-sample predictive ability at the 1-quarter horizon, even though there is significant evidence of in-sample predictive ability. Overall, while there are some discrepancies between the in-sample and out-of-sample test results in Tables 6 and 7, there are nevertheless several financial variables that evince significant in-sample and out-of-sample predictive ability with respect to stock returns over the 1990:2-2000:4 out-of-sample period.

Finally, we control for data mining in our quarterly data by generating critical values for the maximal in-sample and out-of-sample statistics using the data-mining bootstrap procedure. The results are reported in Table 8. At the 1-quarter horizon, the maximal in-sample *t*-statistic for S&P 500 returns in Table 6 is 4.26, corresponding to *cay*. According to the critical values in Panel A of Table 8, the maximal *t*-

¹⁷ Lettau and Ludvigson (2001) find that *cay* has significant predictive ability at the 1-quarter horizon for S&P 500 returns over the 1968:2-1998:3 out-of-sample period according to the *MSE-F* and *ENC-NEW* statistics. Our results in Table 6 show that *cay* also has significant out-of-sample predictive ability over an out-of-sample period beginning in 1990:2, which is similar to the beginning of the out-of-sample period used in Bossaerts and Hillion (1999). Lettau and Ludvigson (2001) is the only extant study we are aware of that employs the *MSE-F* and *ENC-NEW* out-of-sample statistics. Even though the out-of-sample forecasts use estimates of the cointegrating relationship between consumption and wealth based on the full sample of data, Lettau and Ludvigson (2002b, p. 6) argue that “look-ahead” bias is not a problem, as the “parameters in *cay* represent steady-state wealth shares, which, if the theory is true, are clearly known to the agent in equilibrium.”

statistic is significant at the 1% level. According to the same critical values, the t -statistic for the short-term interest rate is also significant at the 1% level at the 1-quarter horizon, while it is significant for the term spread at the 10% level. The maximal out-of-sample *ENC-NEW* statistic at the 1-quarter horizon in Table 6 is 3.22, corresponding to *cay*, and it is significant at the 5% level according to the data-mining robust critical values in Panel A of Table 8. The maximal t -statistics at the 8-quarter and 16-quarter horizons, both corresponding to *cay*, as well as the maximal *MSE-F* and *ENC-NEW* statistics, are not significant at conventional significance levels according to the robust critical values in Panel A of Table 8. The evidence of the predictability of S&P 500 returns at the 1-quarter horizon in Table 6 does not appear to be the result of data mining, but it is more difficult to rule out data mining in Table 6 as far as predictability at longer horizons is concerned.

For CRSP equal-weighted returns, we see from Table 7 that the maximal t -statistic at the 1-quarter horizon is 3.79, corresponding to both *cay* and the short-term interest rate. According to the data-mining robust critical values in Panel B of Table 8, the maximal t -statistic is significant at the 1% level.¹⁸ The maximal out-of-sample *ENC-NEW* statistic in Table 7 is 3.18, corresponding to *cay*, and it is significant at the 5% level according to the robust critical values in Panel B of Table 8. None of the in-sample or out-of-sample maximal statistics are significant at the 8-quarter and 16-quarter horizons according to the robust critical values in Panel B of Table 8. Similar to the results for S&P 500 returns, the evidence of the predictability of CRSP equal-weighted returns at the 1-quarter horizon in Table 7 does not appear to be the result of data mining, but data mining could explain the evidence of predictability at longer horizons in Table 7.

¹⁸ Using the Foster, Smith, and Whaley (1997) distribution that controls for data mining (discussed in footnote 15 above), the maximal R^2 statistics at the 1-quarter horizon of 0.09 in Table 6 and 0.07 in Table 7 are both significant at the 1% level (the 1% critical value is 0.055).

4. Conclusion

In the present paper, we show that there is not a great deal of discrepancy between in-sample and out-of-sample tests of stock return predictability, once we use relatively powerful out-of-sample tests. The extant literature on stock return predictability thus appears to overstate the degree of disparity between in-sample and out-of-sample test results. According to the conventional wisdom, the fact that we identify a number of financial variables with both in-sample and out-of-sample predictive ability suggests that the evidence of stock return predictability is not spurious, as out-of-sample tests help to guard against data mining. However, Inoue and Kilian (2002) recently challenge the conventional wisdom and show that in-sample and out-of-sample tests of predictability are equally susceptible to data mining. They emphasize that the key to controlling for data mining in both in-sample and out-of-sample tests of predictability is the use of appropriate critical values. With this in mind, we test for in-sample and out-of-sample predictability in a data-mining environment suggested by Inoue and Kilian (2002). Using a bootstrap procedure that delivers data-mining robust critical values, we find that certain financial variables, such as the equity share, Fed q , and consumption-wealth ratio, have significant in-sample and out-of-sample predictive ability. Given that we obtain evidence of predictability in an econometric environment that controls for data mining—in addition to potential biases associated with regressor endogeneity and overlapping observations—our in-sample and out-of-sample results strengthen the case for a predictable component in stock returns.¹⁹

¹⁹ Cremers (2001) and Avramov (2002) select the “best” model for predicting returns considering a large number of potential predictors using Bayesian methods. They find some evidence of out-of-sample predictive ability for the best Bayesian models.

Data Appendix

20th Century Annual Data

S&P 500 log real return. Real returns are calculated using the S&P 500 nominal stock price index for January of each year and the annual S&P 500 nominal dividend index available from Robert Shiller's web page at <http://www.econ.yale.edu/~shiller>. The nominal return for year t is calculated as the sum of the nominal stock price in January of year $t + 1$ and dividends for year t divided by the nominal stock price in January of year t . The real return is the nominal return divided by the ratio of the consumer price index in January of year $t + 1$ to the consumer price index in January of year t . We compute the natural logarithm and multiply by 100.

CRSP equal-weighted log real return. The real return for year t is the Center for Research on Security Prices (CRSP) equal-weighted annual nominal return index with dividends for year t divided by the ratio of the consumer price index in January of year $t + 1$ to the consumer price index in January of year t . We compute the natural logarithm and multiply by 100.

Dividend-price ratio, log-level. We calculate the dividend-price ratio corresponding to S&P 500 and CRSP equal-weighted returns. The dividend-price ratio corresponding to S&P 500 returns is based on the Shiller data. The dividend-price ratio in year t is the ratio of nominal dividends during year t to the nominal stock price in January of year $t + 1$. We construct a dividend-price ratio corresponding to CRSP equal-weighted returns by first backing out nominal end-of-year stock price and annual nominal dividend indexes from the annual CRSP equal-weighted nominal returns with and without dividends. The dividend-price ratio in year t is the ratio of the dividends index for year t to the nominal stock price index for the end of year t . For both annual dividend-price ratios, we compute the natural logarithm.

Price-earnings ratio, log-level. The price-earnings ratio in year t is the ratio of the S&P 500 nominal stock price in January of year $t + 1$ to the annual S&P 500 nominal earnings index in year t . The annual S&P 500 nominal earnings index is available from Robert Shiller's web page. We compute the natural logarithm.

Book-to-market ratio, log-level. This is the ratio of the annual Dow Jones Industrial Average (DJIA) index in year t to the annual book value of the DJIA in year t . The data are from the Value Line publication, “A Long-Term Perspective.” We compute the natural logarithm.

Fed q, log-level. This is the ratio of the total market value of equities outstanding to corporate net worth. Robertson and Wright (2000) obtain this data from the Federal Reserve. This ratio is available from Stephen Wright’s web page at <http://www.econ.bbk.ac.uk/faculty/wright> and is used in Robertson and Wright (2000). They note that the Federal Reserve does not refer to this series as “Fed q,” and, while it does not correspond exactly with Tobin’s q , the statistical properties of the two series are very similar. More specifically, the definition for Tobin’s q used in most academic studies is $(S \cdot E + L) / K$, while the Fed q is $S \cdot E / (K - L)$, where S is the price of stocks, E is the number of shares, L is the market value of liabilities, and K is corporate assets. Robertson and Wright (2000) note that the two definitions have the same properties in terms of the sign of the derivative with respect to each of the components. We compute the natural logarithm.

Payout ratio, log-level. This is the ratio of S&P 500 nominal dividends to S&P 500 nominal earnings. We compute the natural logarithm.

Term spread. This is the difference between long-term and short-term government bond yields. The long-term government bond yield data are from Ibbotson Associates (2001, Table A9) and are based on a maturity of approximately 20 years. The short-term yield is the 3-month Treasury bill rate (secondary market) and is from the Federal Reserve Economic Database available from the Federal Reserve Bank of St. Louis web page at <http://www.stls.frb.org/fred/>. The original Treasury bill rates are monthly. We obtain annual observations by averaging over the twelve months comprising a calendar year. As the Treasury bill rate data begin in 1934, we use the short-term Treasury certificate yield data from Homer (1963, Table 49) for the 1927-1933 period.

Default spread. This is the difference between the Moody’s seasoned Baa corporate bond yield and the Moody’s seasoned Aaa corporate bond yield. The corporate bond yield data are also from FRED. The

original corporate bond yield data are monthly. We obtain annual observations by averaging over the twelve months comprising a calendar year.

Short-term interest rate. This is the short-term government bond yield described above.

Equity share. This is the annual total of equity issues (common and preferred) divided by the sum of the annual total of equity issues and the annual total of long-term debt issues (public and private). The data are from Baker and Wurgler (2000), which they compile from the *Federal Reserve Bulletin*. The data are available from Jeffrey Wurgler's web page at <http://www.stern.nyu.edu/~jwurgler/>.

Postwar Quarterly Data

S&P 500 log real return. The nominal return for quarter t is the sum of the nominal stock price at the end of quarter t and the flow of dividends over quarter t and the previous three quarters (divided by four) divided by the nominal stock price at the end of the quarter $t-1$. The real return is the nominal return divided by the ratio of the consumer price index in the first month of quarter $t+1$ to the consumer price index in the first month of quarter t . We compute the natural logarithm and multiply by 400.

CRSP equal-weighted log real return. The real return for quarter t is the CRSP equal-weighted annual nominal return index with dividends for quarter t divided by the ratio of the consumer price index in the first month of quarter $t+1$ to the consumer price index in the first month of quarter t . We compute the natural logarithm and multiply by 400.

Dividend-price ratio, log-level. For the S&P 500 index, the dividend-price ratio in quarter t is the flow of nominal dividends during quarter t and the three previous quarters to the nominal stock price at the end of quarter t . We construct a dividend-price ratio corresponding to CRSP equal-weighted returns by first backing out nominal end-of-quarter stock price and nominal dividend indexes from quarterly CRSP equal-weighted nominal returns with and without dividends. The dividend-price ratio in quarter t is the ratio of the dividends index for quarter t and the previous three quarters to the nominal stock price index for the end of quarter t . For both quarterly dividend-price ratios, we compute the natural logarithm.

Price-earnings ratio, log-level. The price-earnings ratio in quarter t is the ratio of the S&P 500 nominal stock price index at the end of quarter t to the flow of nominal earnings over quarter t and the previous three quarters. We compute the natural logarithm.

Fed q , log-level. This is the ratio of the total market value of equities outstanding to corporate net worth. The quarterly data are available from <http://www.valuingwallstreet.com/updates.shtml>. We compute the natural logarithm.

Payout ratio, log-level. This is the ratio of the S&P 500 nominal dividends index to the S&P 500 nominal earnings index. We compute the natural logarithm.

cay. This is calculated as $cay = c - 0.2985 \cdot a - 0.597 \cdot y$, where c is real per-capita consumption of nondurables and services, a is financial wealth, and y is labor income, all measured in natural logarithms. The coefficients are based on the estimated cointegrating vector in Lettau and Ludvigson (2002). The data are available from Martin Lettau's web page at <http://www.stern.nyu.edu/~mlettau.htm>.

Term spread. This is the difference between long-term and short-term government bond yields. The long-term government bond yield is the 10-year Treasury bond yield, and the short-term government bond yield is the 3-month Treasury bill rate (secondary market). The data are from FRED. The original data are monthly. We obtain quarterly observations by averaging over the three months comprising a quarter.

Default spread. This is the difference between the Moody's seasoned Baa corporate bond yield and the Moody's seasoned Aaa corporate bond yield. The corporate bond yield data are from FRED. The original corporate bond yield data are monthly. We obtain quarterly observations by averaging over the three months comprising a quarter.

Short-term interest rate. This is the quarterly short-term government bond yield described above. Following the convention in the literature, we express the quarterly short-term interest rate in deviations from a four-quarter moving average consisting of the current and three previous quarters.

References

- Ang, Andrew and Geert Bekaert, 2001. Stock return predictability: Is it there? Working Paper No. 8207, National Bureau of Economic Research.
- Avramov, Doron, 2002. Stock return predictability and model uncertainty. *Journal of Financial Economics*, forthcoming.
- Baker, Malcolm and Jeffrey Wurgler, 2000. The equity share in new issues and aggregate stock returns. *Journal of Finance* 55, 2219-2257.
- Bossaerts, Peter and Pierre Hillion, 1999. Implementing statistical criteria to select return forecasting models: What do we learn? *Review of Financial Studies* 12, 405-428.
- Campbell, John Y., 1987. Stock returns and the term structure. *Journal of Financial Economics* 18, 373-399.
- Campbell, John Y., 2000. Asset pricing at the millennium. *Journal of Finance* 55, 1515-1567.
- Campbell, John Y. and Robert J. Shiller, 1988a. The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* 1, 195-228.
- Campbell, John Y. and Robert J. Shiller, 1988b. Stock prices, earnings, and expected dividends. *Journal of Finance* 43, 661-676.
- Campbell, John Y. and Robert J. Shiller, 1998. Valuation ratios and the long-run stock market outlook. *Journal of Portfolio Management*, Winter, 11-26.
- Clark, Todd E. and Michael W. McCracken, 2001a. Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics* 105, 85-110.
- Clark, Todd E. and Michael W. McCracken, 2001b. Evaluating long-horizon forecasts. Manuscript, University of Missouri at Columbia.
- Clements, Michael P. and David F. Hendry, 1998. *Forecasting Economic Time Series*. Cambridge University Press, Cambridge, U.K.
- Cooper, Michael, Roberto C. Gutierrez Jr., and William Marcum, 2001. On the predictability of stock returns in real time. *Journal of Business*, forthcoming.
- Cremers, K.J. Martijn, 2001. Stock return predictability: A Bayesian model selection perspective. *Review of Financial Studies*, forthcoming.
- Diebold, Francis X. and Robert S. Mariano, 1995. Comparing predictive accuracy. *Journal of Economics and Business Statistics* 13, 253-263.
- Fama, Eugene F., 1991. Efficient capital markets: II. *Journal of Finance* 46, 1575-1617.
- Fama, Eugene F. and Kenneth R. French, 1988. Dividend yields and expected stock returns. *Journal of Financial Economics* 22, 3-25.

Fama, Eugene F. and Kenneth R. French, 1989. Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics* 25, 23-49.

Foster, F. Douglas, Tom Smith, and Robert E. Whaley, 1997. Assessing goodness-of-fit of asset pricing models: The distribution of the maximal R^2 . *Journal of Finance* 53, 591-607.

Goetzmann, William N. and Philippe Jorion, 1993. Testing the predictive power of dividend yields. *Journal of Finance* 48, 663-679.

Goyal, Amit and Ivo Welch, 2002. Predicting the equity premium with dividend ratios. Manuscript, University of California at Los Angeles.

Harvey, David I., Stephen J. Leybourne, and Paul Newbold, 1998. Tests for forecast encompassing. *Journal of Business and Economic Statistics* 16, 254-259.

Hodrick, Robert J., 1992. Dividend yields and expected stock returns: Alternative procedures for inference and measurement. *Review of Financial Studies* 5, 357-386.

Homer, Sidney, 1963. *A History of Interest Rates*. Rutgers University Press, New Brunswick, N.J.

Ibbotson Associates, 2001. *Yearbook 2001: Stocks, Bonds, Bills, and Inflation*. Ibbotson Associates, Chicago, Ill.

Inoue, Atsushi and Lutz Kilian, 2002. In-sample or out-of-sample tests of predictability: Which one should we use? Manuscript, University of Michigan.

Kendall, Maurice, 1954. Note on bias in the estimation of autocorrelation. *Biometrika* 41, 403-404.

Kilian, Lutz, 1999. Exchange rates and monetary fundamentals: What do we learn from long-horizon regressions? *Journal of Applied Econometrics* 14, 491-510.

Kirby, Chris, 1997. Measuring the predictable variation in stock and bond returns. *Review of Financial Studies* 10, 579-630.

Kothari, S.P. and Jay Shanken, 1997. Book-to-market, dividend yield, and expected market returns: A time series analysis. *Journal of Financial Economics* 44, 169-203.

Lamont, Owen, 1998. Earnings and expected returns. *Journal of Finance* 53, 1563-1587.

Lettau, Martin and Sydney Ludvigson, 2001. Consumption, aggregate wealth, and expected stock returns. *Journal of Finance* 56, 815-849.

Lettau, Martin and Sydney Ludvigson, 2002a. Understanding trend and cycle in asset values: Bulls, bears, and the wealth effect on consumption. Manuscript, New York University.

Lettau, Martin and Sydney Ludvigson, 2002b. *tay's* as good as *cay*: Reply. Manuscript, New York University.

- Lo, Andrew W. and A. Craig MacKinlay, 1990. Data-snooping biases in tests of financial asset pricing models. *Review of Financial Studies* 3, 431-467.
- Mankiw, N. Gregory and Matthew D. Shapiro, 1986. Do we reject too often? *Economic Letters* 20, 139-145.
- Mark, Nelson C., 1995. Exchange rates and fundamentals: Evidence on long-horizon predictability. *American Economic Review* 85, 201-218.
- McCracken, Michael W., 2000. Asymptotics for out-of-sample tests of causality. Manuscript, University of Missouri at Columbia.
- Nelson, Charles R. and Myung J. Kim, 1993. Predictable stock returns: The role of small sample bias. *Journal of Finance* 48, 641-661.
- Newey, Whitney and Kenneth J. West, 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 10, 129-162.
- Pesaran, M. Hashem and Allan Timmermann, 1995. Predictability of stock returns: Robustness and Economic Significance. *Journal of Finance* 50, 1201-1228.
- Pontiff, Jeffrey and Lawrence D. Schall, 1998. Book-to-market ratios as predictors of market returns. *Journal Financial Economics* 49, 141-160.
- Richardson, Matthew and James H. Stock, 1989. Drawing inferences from statistics based on multiyear asset returns. *Journal of Financial Economics* 25, 323-348.
- Robertson, Donald and Stephen Wright, 2000. The good news and the bad news about long-run stock returns. Manuscript, University of Cambridge.
- Rozeff, Michael, 1984. Dividend yields are equity risk premiums. *Journal of Portfolio Management* 11, 68-75.
- Shaman, Paul and Robert A. Stine, 1988. The bias of autoregressive coefficient estimators. *Journal of the American Statistical Association* 83, 842-848.
- Smithers, Andrew and Stephen Wright, 2000. *Valuing Wall Street* (New York: McGraw-Hill).
- Stambaugh, Robert F., 1986. Biases in regressions with lagged stochastic regressors. Working Paper No. 156, Graduate School of Business, University of Chicago.
- Stambaugh, Robert F., 1999. Predictive regressions. *Journal of Financial Economics* 54, 375-421.
- West, Kenneth D., 1996. Asymptotic inference about predictive ability. *Econometrica* 64, 1067-1084.

Table 1: Descriptive statistics, annual data, 1927-1999

Variable	Mean	Standard deviation	AC(1)	AC(2)	AC(3)	AC(4)
S&P 500 log real return	7.40	18.55	0.04	-0.18	0.05	-0.11
CRSP equal-weighted log real return	9.59	27.67	0.06	-0.21	-0.11	-0.13
Dividend-price ratio, log-level, S&P 500 index	-3.24	0.38	0.79	0.59	0.50	0.39
Dividend-price ratio, log-level, CRSP equal-weighted index	-3.49	0.45	0.90	0.78	0.68	0.64
Price-earnings ratio, log-level	2.65	0.36	0.74	0.54	0.39	0.29
Book-to-market ratio, log-level	-0.49	0.42	0.84	0.65	0.49	0.38
Fed q, log-level	-0.02	0.35	0.82	0.64	0.52	0.44
Payout ratio, log-level	-0.58	0.26	0.81	0.61	0.45	0.36
Term spread	1.45	1.11	0.51	0.19	0.00	-0.04
Default spread	1.14	0.71	0.84	0.62	0.46	0.37
Short-term interest rate	3.88	3.13	0.93	0.83	0.77	0.73
Equity share	0.21	0.11	0.45	0.15	0.04	0.21

Note: The first-order through fourth-order autocorrelations are reported in the fourth through seventh columns.

Table 2: In-sample predictive regression model estimation results and out-of-sample forecasting test results for horizons of 1, 5, and 10 years, annual S&P 500 log real returns

Statistic	Horizon:			Horizon:			Horizon:		
	1-year	5-year	10-year	1-year	5-year	10-year	1-year	5-year	10-year
	<u>Dividend-price ratio, log-level</u>			<u>Payout ratio, log-level</u>			<u>Equity share</u>		
<i>B</i>	0.99	16.48	32.70	-1.57	0.24	-13.14	6.22	9.21	5.62
<i>t</i> -statistic	0.42 [0.73]	1.62 [0.45]	1.89 [0.50]	-0.70 [0.76]	0.05[0.50]	-2.20[0.90]	3.01 [0.00]	1.56[0.12]	1.31[0.21]
<i>R</i> ²	0.00	0.12	0.23	0.01	0.00	0.06	0.11	0.06	0.01
Theil's <i>U</i>	1.03	1.03	0.89	1.02	1.04	1.13	0.96	1.00	1.00
<i>MSE-F</i>	-2.13 [0.73]	-2.05 [0.42]	7.46 [0.18]	-1.44 [0.68]	-2.32[0.51]	-5.97[0.69]	3.50 [0.02]	0.34[0.25]	-0.10[0.35]
<i>ENC-NEW</i>	0.03 [0.47]	3.28 [0.25]	5.33 [0.28]	-0.64 [0.83]	-1.06[0.69]	-2.28[0.80]	3.81 [0.01]	0.66[0.27]	0.02[0.42]
	<u>Price-earnings ratio, log-level</u>			<u>Term spread</u>					
<i>B</i>	2.12	12.19	36.11	0.16	11.10	10.39			
<i>t</i> -statistic	0.94 [0.28]	1.73 [0.25]	5.64 [0.02]	0.07 [0.46]	1.57 [0.09]	1.53[0.14]			
<i>R</i> ²	0.01	0.09	0.39	0.00	0.09	0.04			
Theil's <i>U</i>	1.04	1.06	0.81	1.01	0.95	1.00			
<i>MSE-F</i>	-2.44 [0.87]	-3.60 [0.64]	14.07 [0.04]	-0.81 [0.56]	3.72 [0.04]	0.22[0.29]			
<i>ENC-NEW</i>	0.08 [0.35]	2.03 [0.23]	12.84 [0.04]	-0.39 [0.72]	3.04 [0.06]	0.48[0.29]			
	<u>Book-to-market ratio, log-level</u>			<u>Default spread</u>					
<i>B</i>	1.92	8.25	34.33	0.77	0.94	-2.68			
<i>t</i> -statistic	0.81 [0.41]	0.65 [0.55]	2.55 [0.28]	0.35 [0.41]	0.15[0.52]	-0.24[0.63]			
<i>R</i> ²	0.01	0.03	0.20	0.00	0.00	0.00			
Theil's <i>U</i>	1.22	1.25	0.98	1.00	1.02	1.13			
<i>MSE-F</i>	-11.81 [1.00]	-11.67[.84]	1.34 [0.26]	-0.14 [0.25]	-1.24[0.38]	-5.96[0.74]			
<i>ENC-NEW</i>	-1.76 [0.99]	-2.19 [0.78]	3.79[0.27]	-0.02 [0.38]	-0.60[0.56]	-2.68[0.87]			
	<u>Fed q, log-level</u>			<u>Short-term interest rate</u>					
<i>B</i>	4.29	24.08	53.35	-0.51	0.24	-1.79			
<i>t</i> -statistic	1.90 [0.21]	2.92[0.26]	7.27 [0.06]	-0.23 [0.53]	0.04[0.63]	-0.14[0.57]			
<i>R</i> ²	0.05	0.30	0.72	0.00	0.00	0.00			
Theil's <i>U</i>	1.11	1.06	0.58	1.05	0.00	1.71			
<i>MSE-F</i>	-6.64 [0.97]	-3.30[0.49]	53.24 [0.02]	-3.11 [0.87]	-14.39[0.93]	-17.78[0.90]			
<i>ENC-NEW</i>	1.48 [0.18]	7.12[0.18]	44.29 [0.03]	-0.56 [0.73]	-0.55[0.49]	-2.14[0.70]			

Notes: p -values, computed using the bootstrap procedure described in Section 2.3, are given in brackets. Statistics in bold indicate significance at the 10% level using the p -values given in brackets. 0.00 indicates < 0.005 . $\hat{\beta}$, t -statistic, and R^2 are the OLS estimate of β , its corresponding t -statistic, and the goodness-of-fit measure, respectively, for the in-sample predictive regression model, equation (1), and the horizon indicated. (As noted in the test, we first divide each financial variable by its standard deviation before it enters equation (1), and we take the negative of the price-earnings ratio, log-level, Fed q, log-level, short-term interest rate, and equity share before they enter equation (1).) Theil's U is the ratio of the root-mean-squared forecast errors for the unrestricted and restricted out-of-sample predictive regression models in Section 2.2. The $MSE-F$ statistic is used to test the null hypothesis that the MSE for the unrestricted out-of-sample predictive regression model forecasts is less than the MSE for the restricted out-of-sample predictive regression model forecasts. The $ENC-NEW$ statistic is used to test the null hypothesis that restricted out-of-sample predictive regression model forecasts encompass the unrestricted out-of-sample predictive regression model forecasts.

Table 3: In-sample predictive regression model estimation results and out-of-sample forecasting test results for horizons of 1, 5, and 10 years, annual CRSP equal-weighted log real returns

Statistic	Horizon:			Horizon:			Horizon:		
	1-year	5-year	10-year	1-year	5-year	10-year	1-year	5-year	10-year
	<u>Dividend-price ratio, log-level</u>			<u>Payout ratio, log-level</u>			<u>Equity share</u>		
<i>B</i>	4.30	18.32	24.36	2.09	9.29	7.85	11.67	13.24	6.46
<i>t</i> -statistic	1.30 [0.11]	2.40 [0.07]	2.68[0.11]	0.62 [0.25]	0.91[0.25]	1.34[0.24]	3.89 [0.00]	3.43 [0.01]	1.58[0.15]
<i>R</i> ²	0.02	0.11	0.22	0.01	0.03	0.03	0.18	0.07	0.02
Theil's <i>U</i>	1.01	1.02	0.95	1.02	1.06	0.99	0.95	0.97	1.01
<i>MSE-F</i>	-0.62 [0.13]	-1.05[0.11]	2.70 [0.09]	-1.19 [0.60]	-3.70[0.64]	0.65[0.28]	4.03 [0.01]	2.28 [0.06]	-0.53[0.45]
<i>ENC-NEW</i>	1.58 [0.12]	4.5[0.16]	5.15[0.24]	-0.20 [0.48]	-0.40[0.49]	0.81[0.36]	5.57 [0.00]	1.76 [0.10]	-0.20[0.56]
	<u>Price-earnings ratio, log-level</u>			<u>Term spread</u>					
<i>B</i>	4.23	16.95	19.67	2.55	10.86	13.49			
<i>t</i> -statistic	1.26 [0.17]	2.29[0.14]	1.95[0.27]	0.78 [0.21]	1.42[0.11]	2.35 [0.05]			
<i>R</i> ²	0.02	0.09	0.15	0.01	0.05	0.08			
Theil's <i>U</i>	1.00	0.99	0.98	1.05	1.02	0.97			
<i>MSE-F</i>	-0.14 [0.25]	0.79[0.20]	0.97[0.24]	-3.23 [0.96]	-1.11[0.50]	1.66[0.16]			
<i>ENC-NEW</i>	1.14 [0.10]	3.03[0.14]	2.18[0.24]	-0.93 [0.95]	1.62[0.13]	2.15[0.13]			
	<u>Book-to-market ratio, log-level</u>			<u>Default spread</u>					
<i>B</i>	8.33	16.09	24.15	6.51	12.69	12.90			
<i>t</i> -statistic	2.43 [0.03]	1.91[0.22]	1.68[0.37]	2.03 [0.03]	1.84[0.14]	1.70[0.22]			
<i>R</i> ²	0.08	0.06	0.13	0.06	0.06	0.09			
Theil's <i>U</i>	1.23	1.05	1.01	1.01	1.01	0.98			
<i>MSE-F</i>	-12.21 [1.00]	-3.13[0.44]	-0.69[0.34]	-0.69 [0.47]	-0.34[0.29]	1.24[0.22]			
<i>ENC-NEW</i>	1.16 [0.12]	1.33[0.28]	1.48[0.35]	0.34 [0.24]	0.42[0.33]	0.87[0.31]			
	<u>Fed q, log-level</u>			<u>Short-term interest rate</u>					
<i>B</i>	9.32	24.96	29.65	1.30	9.05	14.03			
<i>t</i> -statistic	2.81 [0.03]	2.68[0.22]	2.37[0.44]	0.40 [0.27]	1.56[0.12]	2.94 [0.05]			
<i>R</i> ²	0.10	0.17	0.29	0.00	0.03	0.11			
Theil's <i>U</i>	1.07	0.97	0.94	1.08	1.23	1.37			
<i>MSE-F</i>	-4.69 [0.92]	1.88[0.23]	3.66[0.26]	-4.96 [0.97]	-10.94[0.84]	-12.51[0.76]			
<i>ENC-NEW</i>	3.52 [0.04]	5.23[0.22]	4.27[0.36]	-0.88 [0.88]	0.46[0.33]	4.68[0.21]			

Notes: p -values, computed using the bootstrap procedure described in Section 2.3, are given in brackets. Statistics in bold indicate significance at the 10% level using the p -values given in brackets. 0.00 indicates < 0.005 . $\hat{\beta}$, t -statistic, and R^2 are the OLS estimate of β , its corresponding t -statistic, and the goodness-of-fit measure, respectively, for the in-sample predictive regression model, equation (1), and the horizon indicated. (As noted in the test, we first divide each financial variable by its standard deviation before it enters equation (1), and we take the negative of the price-earnings ratio, log-level, Fed q, log-level, short-term interest rate, and equity share before they enter equation (1).) Theil's U is the ratio of the root-mean-squared forecast errors for the unrestricted and restricted out-of-sample predictive regression models in Section 2.2. The $MSE-F$ statistic is used to test the null hypothesis that the MSE for the unrestricted out-of-sample predictive regression model forecasts is less than the MSE for the restricted out-of-sample predictive regression model forecasts. The $ENC-NEW$ statistic is used to test the null hypothesis that restricted out-of-sample predictive regression model forecasts encompass the unrestricted out-of-sample predictive regression model forecasts.

Table 4: Data-mining bootstrap critical values, annual S&P 500 and CRSP equal-weighted log real returns

Statistic	1-year horizon			5-year horizon			10-year horizon		
	10%	5%	1%	10%	5%	1%	10%	5%	1%
<u>A. S&P 500 log real returns</u>									
maximal t -statistic	2.61	2.88	3.51	4.64	5.57	7.58	7.56	8.76	11.87
maximal $MSE-F$	3.44	4.39	8.83	17.05	23.44	45.30	36.60	52.21	105.06
maximal $ENC-NEW$	3.89	5.22	7.81	18.78	25.79	42.36	39.30	57.28	102.05
<u>B. CRSP equal-weighted log real returns</u>									
maximal t -statistic	2.50	2.80	3.41	4.33	5.03	6.84	6.29	7.78	11.95
maximal $MSE-F$	3.25	4.41	7.91	16.15	24.56	45.48	30.65	48.82	101.02
maximal $ENC-NEW$	4.15	5.24	7.99	19.62	27.48	49.80	35.85	56.37	96.26

Notes to Table 4: Critical values were computed using the data-mining bootstrap procedure described in Section 2.4. The critical values correspond to the maximum values of the statistics reported in Tables 2 (S&P 500 log real returns) and 3 (CRSP equal-weighted log real returns).

Table 5: Descriptive statistics, quarterly data, 1953:2-2000:4

Variable	Mean	Standard deviation	AC(1)	AC(2)	AC(3)	AC(4)
S&P 500 log real return	7.82	31.43	0.11	-0.07	0.00	0.00
CRSP equal-weighted log real return	8.78	45.21	0.02	-0.06	-0.08	0.06
Dividend-price ratio, log-level, S&P 500 index	-3.41	0.35	0.95	0.89	0.83	0.77
Dividend-price ratio, log-level, CRSP equal-weighted index	-3.74	0.38	0.94	0.89	0.83	0.80
Price-earnings ratio, log-level	2.71	0.36	0.96	0.89	0.83	0.77
Fed q, log-level	-0.27	0.39	0.96	0.91	0.87	0.82
Payout ratio, log-level	-0.70	0.17	0.96	0.87	0.77	0.67
<i>cay</i>	0.61	0.01	0.83	0.68	0.55	0.45
Term spread	1.29	1.12	0.87	0.70	0.59	0.47
Default spread	0.94	0.43	0.93	0.84	0.77	0.70
Short-term interest rate	0.05	1.12	0.71	0.35	0.19	0.06

Notes: The first-order through fourth-order autocorrelations are reported in the fourth through seventh columns.

Table 6: In-sample predictive regression model estimation results and out-of-sample forecasting test results for horizons of 1, 8, and 16 quarters, quarterly S&P 500 log real returns

Statistic	Horizon:			Horizon:		
	1-quarter	8-quarter	16-quarter	1-quarter	8-quarter	16-quarter
	<u>Dividend-price ratio, log-level</u>			<u>cay</u>		
β	2.52	12.71	16.46	9.34	43.96	65.34
t -statistic	1.08 [0.16]	0.50[0.35]	0.33[0.42]	4.26 [0.00]	3.56 [0.01]	3.40 [0.02]
R^2	0.01	0.01	0.01	0.09	0.21	0.25
Theil's U	1.11	1.79	1.45	1.08	0.84	0.75
$MSE-F$	-8.14 [1.00]	-24.75[.98]	-14.7 [0.70]	-6.00[1.00]	14.64 [0.01]	22.35 [0.01]
$ENC-NEW$	-2.40 [1.00]	-7.92[1.00]	-5.82[0.93]	3.22 [0.00]	17.50 [0.00]	13.35 [0.02]
	<u>Price-earnings ratio, log-level</u>			<u>Term spread</u>		
β	1.41	10.72	9.44	6.27	12.31	32.30
t -statistic	0.61[0.50]	0.70[0.55]	0.34[0.70]	2.78 [0.01]	1.32[0.14]	2.07 [0.08]
R^2	0.00	0.01	0.01	0.04	0.02	0.08
Theil's U	1.02	1.21	1.20	1.02	0.99	0.92
$MSE-F$	-1.47[0.87]	-11.2[0.94]	-8.67[0.79]	-1.83[0.96]	0.78[0.25]	4.93[0.11]
$ENC-NEW$	-0.52[0.88]	-3.95[0.95]	-3.66[0.87]	0.54[0.12]	0.84[0.25]	2.78[0.15]
	<u>Fed q, log-level</u>			<u>Default spread</u>		
β	3.30	17.97	30.35	3.25	-1.31	6.83
t -statistic	1.44[0.44]	0.84[0.84]	0.81[0.85]	1.43 [0.07]	-0.12[0.50]	0.29[0.42]
R^2	0.01	0.03	0.05	0.01	0.00	0.00
Theil's U	1.07	1.51	1.38	1.02	1.04	1.07
$MSE-F$	-5.51[0.98]	-20.1 [0.96]	-13.3 [0.80]	-1.55[0.92]	-2.55[0.67]	-3.64[0.68]
$ENC-NEW$	-1.63[0.97]	-7.24[1.00]	-5.45[0.96]	-0.61[0.94]	-1.24[0.75]	-1.74[0.76]
	<u>Payout ratio, log-level</u>			<u>Short-term interest rate</u>		
β	2.11	-2.49	-0.43	8.42	12.67	11.63
t -statistic	0.91[0.18]	-0.17[0.53]	-0.02[0.49]	3.82 [0.00]	1.62 [0.07]	1.19[0.15]
R^2	0.00	0.00	0.00	0.07	0.02	0.01
Theil's U	1.05	1.15	1.01	1.02	1.03	1.00
$MSE-F$	-4.21[0.99]	-8.66[0.91]	-0.64[0.52]	-1.53[0.92]	-2.07[0.79]	-0.17[0.47]
$ENC-NEW$	-1.23[0.99]	-3.31[0.94]	-0.23[0.57]	0.33[0.20]	-0.71[0.80]	0.00[0.49]

Notes: p -values, computed using the bootstrap procedure described in Section 2.3, are given in brackets. Statistics in bold indicate significance at the 10% level using the p -values given in brackets. 0.00 indicates < 0.005 . β , t -statistic, and R^2 are the OLS estimate of β , its corresponding t -statistic, and the goodness-of-fit measure, respectively, for the in-sample predictive regression model, equation (1), and the horizon indicated. (As noted in the test, we first divide each financial variable by its standard deviation before it enters equation (1), and we take the negative of the price-earnings ratio, log-level, Fed q, log-level, and short-term interest rate before they enter equation (1).) Theil's U is the ratio of the root-mean-squared forecast errors for the unrestricted and restricted out-of-sample predictive regression models in Section 2.2. The $MSE-F$ statistic is used to test the null hypothesis that the MSE for the unrestricted out-of-sample predictive regression model forecasts is less than the MSE for the restricted out-of-sample predictive regression model forecasts. The $ENC-NEW$ statistic is used to test the null hypothesis that restricted out-of-sample predictive regression model forecasts encompass the unrestricted out-of-sample predictive regression model forecasts.

Table 7: In-sample predictive regression model estimation results and out-of-sample forecasting test results for horizons of 1, 8, and 16 quarters, quarterly CRSP equal-weighted log real returns

Statistic	Horizon:			Horizon:		
	1-quarter	8-quarter	16-quarter	1-quarter	8-quarter	16-quarter
	<u>Dividend-price ratio, log-level</u>			<u>Cay</u>		
β	6.69	39.00	48.00	12.08	45.93	48.41
t -statistic	2.05 [0.09]	2.68[0.13]	1.87[0.35]	3.79 [0.00]	2.94 [0.02]	2.64 [0.05]
R^2	0.02	0.11	0.10	0.07	0.14	0.09
Theil's U	1.04	1.93	2.18	1.04	1.04	0.70
$MSE-F$	-3.12 [0.95]	-26.34 [1.0]	-22.10 [1.0]	-2.84[0.99]	-2.78[0.73]	28.45 [0.01]
$ENC-NEW$	-0.09 [0.60]	-5.22[0.98]	-5.96[0.99]	3.18 [0.00]	17.87 [0.00]	26.42 [0.01]
	<u>Price-earnings ratio, log-level</u>			<u>Term spread</u>		
β	4.38	37.13	51.56	7.12	-4.69	-4.53
t -statistic	1.33[0.18]	2.30[0.14]	1.53[0.34]	2.18 [0.02]	-0.41[0.64]	-0.28[0.59]
R^2	0.01	0.09	0.10	0.03	0.00	0.00
Theil's U	1.02	1.74	1.93	0.99	1.08	1.19
$MSE-F$	-1.93[0.92]	-24.10[1.0]	-20.5 [0.98]	0.91 [0.07]	-4.93[0.85]	-8.18[0.92]
$ENC-NEW$	-0.35[0.82]	-5.38[0.98]	-6.22[1.00]	1.18 [0.03]	-1.98[0.88]	-3.46[0.95]
	<u>Fed q, log-level</u>			<u>Default spread</u>		
β	6.27	38.73	54.81	6.49	0.73	-6.55
t -statistic	1.96[0.23]	1.94[0.39]	1.40[0.66]	1.99 [0.02]	0.06[0.49]	-0.35[0.58]
R^2	0.02	0.09	0.10	0.02	0.00	0.00
Theil's U	1.04	1.67	1.67	1.00	1.01	0.99
$MSE-F$	-3.07[0.92]	-23.15[1.0]	-17.9 [0.91]	0.29[0.16]	-0.52[0.46]	0.65[0.31]
$ENC-NEW$	-0.21[0.72]	-4.23[0.92]	-5.39[0.96]	0.35[0.15]	-0.25[0.55]	0.33[0.38]
	<u>Payout ratio, log-level</u>			<u>Short-term interest rate</u>		
β	3.35	-7.96	-36.14	12.02	9.85	0.31
t -statistic	1.00[0.15]	-0.58[0.69]	-1.11[0.77]	3.79 [0.00]	0.88[0.23]	0.03[0.45]
R^2	0.01	0.00	0.05	0.07	0.01	0.00
Theil's U	1.00	1.16	1.78	0.99	0.99	1.03
$MSE-F$	-0.21[0.46]	-9.34[0.90]	-19.19[1.0]	0.60[0.11]	1.06[0.18]	-1.41[0.75]
$ENC-NEW$	0.17[0.25]	-3.35[0.92]	-4.88[0.97]	1.27 [0.03]	0.76[0.21]	-0.68[0.82]

Notes: p -values, computed using the bootstrap procedure described in Section 2.3, are given in brackets. Statistics in bold indicate significance at the 10% level using the p -values given in brackets. 0.00 indicates < 0.005 . β , t -statistic, and R^2 are the OLS estimate of β , its corresponding t -statistic, and the goodness-of-fit measure, respectively, for the in-sample predictive regression model, equation (1), and the horizon indicated. (As noted in the test, we first divide each financial variable by its standard deviation before it enters equation (1), and we take the negative of the price-earnings ratio, log-level, Fed q, log-level, and short-term interest rate before they enter equation (1).) Theil's U is the ratio of the root-mean-squared forecast errors for the unrestricted and restricted out-of-sample predictive regression models in Section 2.2. The $MSE-F$ statistic is used to test the null hypothesis that the MSE for the unrestricted out-of-sample predictive regression model forecasts is less than the MSE for the restricted out-of-sample predictive regression model forecasts. The $ENC-NEW$ statistic is used to test the null hypothesis that restricted out-of-sample predictive regression model forecasts encompass the unrestricted out-of-sample predictive regression model forecasts.

Table 8: Data-mining bootstrap critical values, quarterly S&P 500 and CRSP equal-weighted log real returns

Statistic	1-year horizon			5-year horizon			10-year horizon		
	10%	5%	1%	10%	5%	1%	10%	5%	1%
<u>A. S&P 500 log real returns</u>									
maximal t -statistic	2.67	3.01	3.50	4.18	4.80	5.92	5.72	6.70	9.72
maximal $MSE-F$	2.80	3.44	4.86	20.83	27.22	43.09	50.92	78.73	163.42
maximal $ENC-NEW$	2.28	2.84	4.21	18.85	27.04	44.07	51.02	74.12	163.32
<u>B. CRSP equal-weighted log real returns</u>									
maximal t -statistic	2.70	2.95	3.57	4.28	4.91	5.88	5.70	6.85	9.13
maximal $MSE-F$	2.68	3.46	5.05	20.03	28.24	46.57	37.15	52.55	107.96
maximal $ENC-NEW$	2.02	2.64	3.89	16.38	22.52	39.49	33.78	51.98	86.74

Notes to Table 8: Critical values were computed using the data-mining bootstrap procedure described in Section 2.4. The critical values correspond to the maximum values of the statistics reported in Tables 6 (S&P 500 log real returns) and 7 (CRSP equal-weighted log real returns).