THE JOURNAL OF FINANCE • VOL. LII, NO. 2 • JUNE 1997

Assessing Goodness-of-Fit of Asset Pricing Models: The Distribution of the Maximal R^2

F. DOUGLAS FOSTER, TOM SMITH, and ROBERT E. WHALEY*

ABSTRACT

The development of asset pricing models that rely on instrumental variables together with the increased availability of easily-accessible economic time-series have renewed interest in predicting security returns. Evaluating the significance of these new research findings, however, is no easy task. Because these asset pricing theory tests are not independent, classical methods of assessing goodness-of-fit are inappropriate. This study investigates the distribution of the maximal R^2 when k of m regressors are used to predict security returns. We provide a simple procedure that adjusts critical R^2 values to account for selecting variables by searching among potential regressors.

PERHAPS THE MOST ELUSIVE goal in finance is the ability to predict security returns. Researchers have devoted no small amount of time and effort to understanding the return generating process. Recent studies include investigations of return anomalies and developments of conditional asset pricing models with time-varying parameters. Some studies conclude that security returns are predictable—a notion that historically would have been treated with skepticism. While interesting, this research raises concerns about whether these insights are real or whether we have simply become too familiar with the available data.

This study adds to the debate about whether security returns are predictable. Our concern is that tests of predictive power are made using classical statistical tests (where the implicit assumption is that only one test is made with a particular data set), yet most research designs rely upon past work. If

* Foster is from The University of Iowa, Smith is from the University of New South Wales, and Whaley is from Duke University. We are grateful for the helpful comments of seminar participants at the Australian Finance Association Conference, University of Arizona, Australian Graduate School of Management, Berkeley Program in Finance, University of British Columbia, University of California at Berkeley, University of Central Florida, University of Iowa, University of Michigan, Monash University, North Carolina State University, the Pacific Basin Finance Conference in Hong Kong, University of Pennsylvania, Queensland University of Technology, Rice University, the Second Conference on Financial Economics and Accounting at SUNY Buffalo, Stanford University, University of Western Australia, the Western Finance Association Annual Meetings in San Francisco, Fischer Black, Louis Chan, Bernard Dumas, Fred Feinberg, Campbell Harvey, Nancy Keeshan, Allan Kleidon, Andrew Lo, Kevin McCardle, Bob Nau, George Oldfield, Matthew Richardson, Jim Smith, René Stulz (the editor), S. Viswanathan, and an anonymous referee. Research support was received from the Business Associates' Fund (Smith and Whaley), Fuqua School of Business, Duke University. This research would not have been possible without the generous support of the North Carolina Super Computing Center.

past work uses similar data (e.g., from a similar time period), questions of predictability become especially nettlesome. While any single researcher is unlikely to have completed an exhaustive search of a large set of potential regressors in selecting an instrument set, the effect may be the same if the researcher knows the results of related studies.[1]

To apply our insights, we consider tests of conditional asset pricing models. By recognizing that model parameters may change through time, conditional asset pricing models allow for more powerful tests. The implementation of these tests, however, requires prediction of security returns using the instrumental variables that form the investors' information set. Since the identity of the instruments is unknown, researchers search available information and select variables that best describe the data. The problem with the goodness-of-fit approach is that it is usually evaluated using classical cutoff levels. If the wrong instrument set (i.e., one that has no true predictive power) is chosen, the conditional asset pricing test is misspecified.

The purpose of this article is to provide a simple procedure to test the null hypothesis that all of the slope coefficients of an Ordinary Least Squares (OLS) regression used to predict returns are equal to zero where we assume that (a) only k of m potential regressors are used, (b) all possible regression combinations are tried, and (c) only the regression with the highest R^2 is reported. The alternative hypothesis is that at least one of the OLS slope coefficients is not equal to zero (i.e., the regression has predictive power) given the same exhaustive variable-selection procedure.

The article proceeds as follows. In Section I, we discuss the variable-selection problem and illustrate how an improper use of classical statistical inference can lead to erroneous conclusions. In Section II, we develop a test procedure. Relying on the applied statistics literature, we start with the distribution of the R^2 statistic under classical OLS assumptions. Next we discuss a bound for the maximal R^2 distribution when the best k of m potential regressors are selected. This bound represents a conservative test for determining whether a model is statistically significant when the researcher has access to m explanatory variables and uses only k of them. We also solve numerically for the maximal R^2 distribution. Finally, we consider an approximation developed by Rencher and Pun (1980). In Section III, we discuss the implications of our work for tests of monthly security return prediction. Section IV contains a summary.

I. The Variable-Selection Problem

Concern about overfitting data in tests of financial models is not new. Merton (1987, p. 207) warns that researchers may find return anomalies because they are too close to the data. Ross (1989) points out that, in searching return data for anomalies, researchers will find patterns that are at odds with the current paradigm. Lo and MacKinlay (1990, 1992) investigate data-snooping biases and point out that grouping stocks into portfolios induces bias in statistical

[1] See, for example, Denton (1985).

tests. Black (1992) discusses a variety of other data-snooping biases and highlights the relative roles of theory and data in understanding how one can estimate expected security returns. Granger and Newbold (1974) show how high R^2 statistics may result from a spurious regression problem.

In this study, we focus on a *single* potential source of inflated R^2 statistics—choosing a subset k of m possible explanatory variables. This potential for overfitting is well known in the applied statistics literature. Freedman (1983, p. 152), for example, states ". . . in a world with a large number of unrelated variables and no clear a priori specification, uncritical use of standard methods will lead to models that appear to have a lot of explanatory power." Following the applied statistics work of Rencher and Pun (1980), Miller (1984, 1990) and Hjorth (1994), we demonstrate that the bias resulting from this form of variable selection can be important.[2]

A. *The OLS Model*

To help formulate the variable-selection problem in an asset pricing theory context, let y be a $t \times 1$ vector of security returns. To predict these returns, a researcher chooses k of m (where $k < m$) potential regressors. In studies predicting monthly security returns, researchers typically choose five or six regressors (k). The number of potential regressors (m), however, is virtually without limit. Among the candidates are stock and bond prices and returns, macroeconomic variables and accounting variables, not only from the domestic economy but from all economies worldwide. In addition, there are limitless possible linear and nonlinear transformations of these variables (e.g., the difference between the long-term and short-term Treasury rates, the logarithm of market capitalization, the square of market return).

The simplest way to choose among the regressors is to rely on the goodness-of-fit as measured by the regression R^2 (or F-statistic). In the context of return predictability, the m potential regressors are predetermined variables representing elements of past information sets. From the m variables, the researcher chooses only k regressors. The regression equation is:

$$y = c + x\beta + \varepsilon, \tag{1}$$

where c is an intercept, x is a $t \times k$ nonstochastic data matrix, β is a $k \times 1$ vector of coefficients and ε is a $t \times 1$ vector of independent disturbance terms each distributed as $N(0, \sigma^2)$. The number of possible model regression specifications is $\binom{m}{k}$.

[2] In our analysis, the number of regressors, k, is fixed. Choosing the number of regressors (i.e., "dimensionality selection") also produces biases and is an interesting problem in its own right. Breiman (1992) points out that there are a number of commonly-used ad hoc methods (e.g., F to enter, F to delete, adjusted R^2, and Mallows C_p) for choosing the submodel dimension. He also introduces a technique known as the "little bootstrap" that gives almost unbiased estimation for submodel prediction errors and uses these for model selection. Other important work on this problem includes Breiman and Specter (1992) and Miller (1990).

B. An Illustration

To illustrate the degree of possible bias in the R^2 from choosing the best k of m regressors, consider the following simulation. First, we generate 250 observations for 51 variables—fifty potential regressors and one dependent variable. The sample size of 250 was chosen to be about the same as those used in monthly return prediction studies. Each generated variable is normally distributed with mean zero and unit variance, and is independent of all the others. With independence, the regression is expected to fit poorly.

Next, using the generated sample, we perform the OLS regression (1). When the dependent variable is regressed on all fifty independent variables, the F-value is 1.172 (with a p-value of 0.22), the R^2 is 0.228, and the $\bar{R}^2$ is 0.034. The five most extreme t-statistics are 2.45, −2.24, 2.02, 1.92, and 1.83. With fifty regressors, five (i.e., ten percent) should have t-statistics whose magnitude is 1.645 or higher through random chance. Based on these results, y and x are unrelated, and hence the design of the data is revealed.

Finally, again using the generated sample, we search across all possible combinations of five regressors to find that set that maximizes R^2. The F-statistic is 4.137 (with a p-value of 0.0013), the R^2 is 0.078, and the $\bar{R}^2$ is 0.059. The t-statistics of the five "best" regressors are 2.58, −2.21, 2.02, 2.68, and 1.73. Absent the knowledge that an exhaustive search had been performed, the relation between y and x would be regarded as significant. Nothing is further from the truth, however. The inference is wrong because the effects of selecting the five "best" regressors from the set of fifty are ignored. We now demonstrate techniques similar to an F-test that can be applied to censored regressions to arrive at the same inference as the full regression.

II. The Distribution of the Maximal R^2

The key to the analysis lies in the distribution of the maximal R^2. To estimate the β coefficients in equation (1), we use their OLS values, $b \equiv (x'x)^{-1} x'y$. With this definition of b, the proportion of the variation in y that is explained by changes in x is the R^2, and the R^2 is distributed as a

$$\text{Beta}\left(\frac{k}{2}, \frac{t-(k+1)}{2}\right)$$

under the null hypothesis that $\beta = 0$.[3] The significance of the regression can be assessed with the R^2 statistic in the same manner as with the F-statistic.[4] The 95 percent cutoff value of the R^2 from the single, five-variable regression is 0.044. Using this cutoff value for the censored regression in Section I, we falsely infer that at least one of the β coefficients is not zero. The incorrect inference arises because the classical testing procedure does not take into

[3] For the distribution of R^2 with a null hypothesis other than $\beta = 0$, see Cramer (1987).

[4] The significance of the regression can also be assessed using the $\bar{R}^2$ statistic since the unadjusted value is simply $R^2 = 1 - \{[t-(k+1)]/t - 1\}(1 - \bar{R}^2)$.

account the amount of data used to select explanatory variables. We now show how to adjust the cutoff value of R^2 assuming the researcher chooses the "best" k of the m (where $k < m$) regressors by maximizing the regression R^2.

A. Bounding the Distribution

To begin, we consider the case where the regressions are independent (i.e., the x matrix and the y vector are independent between regressions). The distribution function of the maximal R^2, $U_{R^2}(\cdot)$, is:

$$U_{R^2}(r) = \Pr(R_1^2 \leq r,\ R_2^2 \leq r,\ \ldots,\ R_{(m,k)}^2 \leq r) = [\text{Beta}(r)]^{(m,k)}, \qquad (2)$$

where $Beta(\cdot)$ is the cumulative distribution function of the beta density function with $k/2$ and $[t - (k + 1)]/2$ degrees of freedom. Expression (2) is the standard order statistic argument that the probability that all R^2s are less than some cutoff is the product of the probabilities that each R^2 is less than the cutoff.[5] Because we have $\binom{m}{k}$ potential regressions, $\binom{m}{k}$ terms appear on the right-hand side of equation (2).

Expression (2) is a relatively straightforward way to compute the distribution of the maximal R^2; however, it cannot be used directly for our purposes. In our regression experiment, the y does not change between regressions, the $\binom{m}{k}$ x matrices have overlapping elements, and the regressors in the x matrix may be correlated.[6] Consequently, we cannot derive a general form for the joint probability distribution of the R^2 and move to a weaker bound of the joint distribution function based on the Bonferroni inequality:

$$U_{R^2}(r) \geq 1 - \left\{[1 - \text{Beta}(r)]\binom{m}{k}\right\}. \qquad (3)$$

Expression (3) states that the upper tail probability of the univariate distribution is scaled by the number of regressions performed. This bound holds for general correlations among the regressions (i.e., both the effect of a shared y vector and the correlations in the x matrix).

To develop a more intuitive understanding of the plausible implications of expressions (2) and (3) for tests of monthly security return prediction, reconsider the illustration in Section I where $m = 50$, $k = 5$, and $t = 250$. The 95 percent cutoff values computed using equations (2) and (3) are both approximately 0.164. Since the R^2 of our "best" five-variable regression is only 0.078, we cannot reject the hypothesis that the β are zero—the same inference as the first regression when all fifty regressors were used.

The cutoff level of equation (3) is a bound and is therefore biased toward not rejecting the null hypothesis. If an R^2 exceeds equation (3), we can be confident

[5] See David (1981).

[6] If the x matrix does not have overlapping elements and there is no correlation among regressors, Kimball (1951) shows that expression (2) is a bound on the distributional of the maximal R^2. Such circumstances would arise, for example, where only one regressor is selected.

that the regression fit is not a result of exhaustive variable selection. If an R^2 is less than equation (3), however, we are uncertain.

With added computational cost, we can be more precise in our testing. Although we do not know the exact distribution of the maximal R^2, it can be computed numerically using Monte Carlo simulation. To illustrate, we use the parameters from the example in Section I. The simulation begins with generating 250 standard normal observations for fifty potential regressors and a dependent variable that is uncorrelated (in population) with all of the regressors. We then select the best five regressors from the fifty potential regressors and record the R^2 from the regression fit. We repeat this exercise 100 times to create a numerical distribution of the maximal R^2. Under this procedure, we find that the 95 percent cutoff R^2 value is 0.117. Since this cutoff value exceeds 0.078, we cannot reject the null hypothesis that all β_i are zero.

The numerically generated cutoff is the an approximation for the exact distribution. Unlike equations (2) and (3), it is not biased towards the null. Moreover, correlation among potential regressors does not appear to affect the numerical distribution of the maximal R^2. In separate simulations, we repeat this analysis for low and high positive correlation among the potential regressors and find similar cutoff values.[7]

B. An Approximation

Rencher and Pun (1980) use extreme value theory to derive an asymptotic distribution of the maximal R^2. They consider the same basic design, and, by assuming independence between $N = \binom{m}{k}$ possible regression combinations, they show that the γ percent cutoff level of the maximum R^2 is

$$R^2_\gamma \approx F^{-1}\left[1 + \frac{\ln(\gamma)}{N}\right], \tag{4}$$

where F^{-1} is the inverse of the Beta cumulative distribution function. Expression (4) is analogous to equation (2).

Rencher and Pun go on to note that the independence assumption is unlikely to apply. To correct for the dependence among possible regression combinations, they adjust the value of N in expression (4). The adjusted N is smaller and reflects the effects of using the same y vector and common columns in the various x matrix combinations. Using simulated data, they choose the best regressors using a step-wise procedure. Working with their numerical results, they fit the functional form of $\ln(N)^{cN^d}$ for N in expression (4). For the cutoff R^2 values, they find $c = 1.8$ and $d = 0.04$ are appropriate. The Rencher and Pun approximation for the cutoff level (dubbed the "rule-of-thumb") is therefore

$$R^2_\gamma \approx F^{-1}\left[1 + \frac{\ln(\gamma)}{\ln(N)^{1.8N^{0.04}}}\right]. \tag{5}$$

[7] This is not surprising since, in going from the uncorrelated to the correlated case, we simply multiply the potential regressors by a constant matrix (the Cholesky half of the variance-covariance matrix).

The rule-of-thumb approximation works well for our illustration. Using expression (5), the 95 percent cutoff for choosing the best five of fifty regressors with 250 data points is 0.119, which is very close to our Monte Carlo value of 0.117.[8] Because the rule-of-thumb is based on numerically generated densities over a fairly narrow parameter space[9] and an assumed functional form, however, one must use it with care.

C. Applying the Techniques

To provide a general sense for the application of these techniques, we compute the Bonferonni bound and the Rencher/Pun rule-of-thumb for parameter ranges typically found in studies of monthly security return prediction. In Table I, for example, we assume that the researcher chooses the "best" five of m potential regressors, where m ranges up to 500. The sample size (t) ranges up to 1000, or approximately eighty years of monthly data. Table II contains the 95 percent cutoff values when the researcher chooses the "best" k of m potential regressors and the sample size t is fixed at 250, or approximately twenty years of monthly data.

The results shown in Table I reveal that, holding sample size constant, the critical R^2 values increase with the number of potential regressors. With a sample size of fifty and with ten potential regressors, the Bonferonni bound is 0.413. In other words, a regression R^2 must exceed 0.413 to be assured that an exhaustive search of all possible regression combinations could not produce a model that does a better job of predicting security returns. If the number of potential regressors goes up to, say, 500, the critical R^2 as determined by the Bonferonni bound is 0.780. The bound grows higher as the number of potential regressors increases because, with each additional regressor, there is an increased chance of finding significant results.

Holding the number of potential regressors constant, Table I shows that the critical R^2 value decreases with sample size. With ten potential regressors and fifty observations, the critical R^2 value as determined by the Bonferonni bound is 0.413. This value decreases to 0.224 at a sample size of 100, 0.094 at a sample size of 250, and so on. As the number of observations increases, the chance that an exhaustive search of all possible regression combinations will produce a model that fits the data well becomes small.

[8] We also compute numerically (using 1,000 repetitions to generate the numerical distributions) the 95 percent cutoff levels for the cases where $m = 10$ and $m = 25$. For ten potential regressors, the numerical R^2 cutoff is 0.067. As reported in Table I, the corresponding rule-of-thumb approximation is 0.079, and the Bonferonni bound is 0.094. For 25 potential regressors, the numerical R^2 cutoff is 0.096, the rule-of-thumb is 0.103, and the Bonferonni bound is 0.136.

[9] Indeed, Rencher and Pun (1980, p. 52) are careful to note that the approximation (5) should only be used for values of k, m, and n bracketed by the parameters used in deriving the functional form's parameters. In their words, "extrapolation beyond may be risky and needs further investigation." The parameter ranges used in their work are as follows: the number of selected regressors, $k = 2, \ldots, 10$; the number of potential regressors, $m = 5, \ldots, 40$; and, the number of observations, $t = 5, \ldots, 60$.

Table I

95 Percent Cutoff Values for the "Best" Five-Variable Regression R-Squared Given Different Sample Sizes (t) and Different Numbers of Potential Regressors (m)

This table reports the 95 percent confidence limit for R^2 for the null hypothesis that all of the slope coefficients of an Ordinary Least Squares (OLS) regression are equal to zero where (a) only five of m potential regressors are used, (b) all possible regression combinations are tried, and (c) only the regression with the highest R^2 is reported. The alternative hypothesis is that at least one of the OLS slope coefficients is not equal to zero. The Bonferonni inequality is a bound and therefore represents a conservative test. The Rencher/Pun (1980) rule-of-thumb is an approximation of the exact distribution.

Number of Potential Regressors (m)	Sample Size (t)				
	50	100	250	500	1000
Panel A: Bonferonni Bound					
10	0.413	0.224	0.094	0.048	0.024
25	0.548	0.314	0.136	0.070	0.036
50	0.621	0.369	0.164	0.085	0.043
100	0.679	0.417	0.189	0.099	0.050
250	0.742	0.474	0.221	0.116	0.060
500	0.780	0.513	0.244	0.129	0.067
Panel B: Rencher/Pun Rule-of-Thumb					
10	0.360	0.191	0.079	0.040	0.020
25	0.444	0.244	0.103	0.052	0.026
50	0.495	0.278	0.119	0.061	0.031
100	0.545	0.312	0.135	0.070	0.035
250	0.610	0.361	0.160	0.083	0.042
500	0.658	0.400	0.180	0.094	0.048

The Rencher/Pun rule-of-thumb results reported in Panel B of Table I are uniformly lower than those of the Bonferonni bound. As noted earlier, the Bonferonni test is conservative. If an R^2 exceeds equation (3), we can be confident that the regression fit is not a result of exhaustive variable selection. There is a region below the Bonferonni cutoff, however, where we are uncertain. The rule-of-thumb approximates the exact distribution of the maximal R^2. The entry corresponding to a sample size of 250 and fifty potential regressors is 0.119, the cutoff value discussed in the illustration of the Rencher/Pun technique.

Table II offers some insights regarding the critical R^2 levels when the number of selected regressors (k) and the number of potential regressors (m) vary while the sample size (t) remains fixed. Holding the number of potential regressors constant, the critical R^2 value increases with the number of selected variables. With each added explanatory variable, the amount of variation explained by the regression increases. Consequently, the threshold for significance also increases.

Table II

95 Percent Cutoff Values for the "Best" k-Variable Regression R-Squared Given Different Numbers of Potential Regressors (m) and a Fixed Sample Size of 250 (i.e., $t = 250$)

This table reports the 95 percent confidence limit for R^2 for the null hypothesis that all of the slope coefficients of an Ordinary Least Squares (OLS) regression are equal to zero where (a) only k of m potential regressors are used, (b) all possible regression combinations are tried, and (c) only the regression with the highest R^2 is reported. The alternative hypothesis is that at least one of the OLS slope coefficients is not equal to zero. The Bonferonni inequality is a bound and therefore represents a conservative test. The Rencher/Pun (1980) rule-of-thumb is an approximation of the exact distribution.

Number of Potential Regressors (m)	Number of Regressors Selected (k)				
	1	2	3	4	5
Panel A: Bonferonni Bound					
10	0.036	0.055	0.071	0.084	0.094
25	0.040	0.068	0.094	0.116	0.136
50	0.044	0.079	0.110	0.138	0.164
100	0.048	0.089	0.126	0.159	0.189
250	0.055	0.102	0.146	0.185	0.221
500	0.059	0.112	0.160	0.204	0.244
Panel B: Rencher/Pun Rule-of-Thumb					
10	0.027	0.046	0.060	0.071	0.079
25	0.032	0.054	0.072	0.088	0.103
50	0.035	0.060	0.081	0.100	0.119
100	0.038	0.066	0.090	0.113	0.135
250	0.042	0.073	0.101	0.130	0.160
500	0.045	0.078	0.110	0.144	0.180

Together, Tables I and II provide a means of gauging the significance of monthly return prediction tests. Generally speaking, past tests use less than forty years of data and have five regressors or less. Where the sample size or the number of potential regressors is not reported in the tables, interpolating between the values reported in the tables will produce reasonably accurate cutoff R^2 values.

III. Evaluation of Apparent Return Predictability

In Section I, we showed the potentially misleading inferences that can result when k of m regressors are chosen, and, in Section II, we showed how the classical cutoff R^2 levels may be adjusted to account for the variable-selection bias. In this section, we discuss two interrelated issues. First, we discuss the difficulties of designing an out-of-sample testing procedures to circumvent the overfitting problem. Second, we use the results of past studies of return predictability to illustrate the application of the Bonferroni bound and the Rencher/Pun rule-of-thumb.

A. Out-of-Sample Predictions

Many investigators believe that the unfortunate consequences of data-mining can be mitigated if not eliminated by out-of-sample tests. Some gain comfort, for example, in knowing that the results extend to other markets internationally, while others examine different sectors of the same market (e.g., different industry and size portfolios). Whether or not this is effective depends, of course, on the extent to which the new data is truly out-of-sample. If returns in the other markets or sectors of the same markets are correlated even slightly with those that have been previously mined, the investigator will derive false confidence from tests of this nature. A simulation experiment illustrates this point.

First, we generate samples of 250 "returns" for "country," "industry," and "size" portfolios. The simulated portfolio returns are normally distributed with mean zero and unit variance. Since the correlation structure between these simulated portfolio returns will drive the results, we set the correlation structure using historical estimates. The country portfolio correlation structure, for example, is based on the estimated correlations from monthly excess U.S. dollar returns from February 1970 through December 1989, and the industry portfolio correlation structure is formed using excess U.S. industry portfolio returns from May 1959 through December 1986.[10] The size portfolio correlation structure is based on monthly excess returns of U.S. stocks from January 1926 through December 1984.

Table III reports the correlation structure used in generating the portfolio returns. The first row of the table contains the base portfolio for each of the three categorizations. The value of 0.48 reported for Australia is the assumed correlation between the returns of the U.S. stock portfolio and the Australian stock index portfolio used in the simulation. The country portfolio return correlations range from 0.13 for Austria to 0.72 for Canada. The industry portfolio return correlations are higher, ranging from 0.41 for textile/trade to 0.72 for finance/real estate. The size portfolio returns are the most highly correlated, decreasing monotonically from 0.95 for the second smallest decile to 0.71 for the largest decile.

The simulated portfolio returns for the United States, the petroleum industry, and the decile of the smallest capitalization firms are used as the base portfolios. For each base case, we then generate return series for fifty regressors. Each regressor is normally distributed with zero mean and unit variance, and is independent of all others (and the base portfolio return series). From the fifty regressors, we choose the "best" five by maximizing the regression R^2. After recording the R^2 for the base portfolio, we use the five identified regressors to compute the R^2 values of the other portfolios in the category. We repeat the procedure 100 times. In Table III, we report the ratio of the average R^2 of each correlated portfolio to the average R^2 of the base portfolio and call this

[10] We are grateful to Campbell Harvey for providing his estimates of the correlation matrices for the industry and country portfolios.

Table III

Variable-Selection Bias for Correlated Portfolios

This table reports the results of simulations designed to deduce the expected explanatory power of a set of factors on the returns of a correlated portfolio given the amount of explanatory power of the same factors on the returns of a base portfolio. Each simulation involves generating return series of length 250 for each of a number of correlated portfolios formed on the basis of "country," "industry," and "size." Although the returns are generated to have zero mean and unit variance, the assumed correlation structures are based on actual correlations computed using monthly historical return data and are reported below. Time series returns for fifty regressors are also generated, assuming each regressor is independent of all others and of the correlated portfolios. Using the base portfolio returns under each portfolio categorization (i.e., "United States" under "Country portfolios," "Petroleum" under "Industry portfolios," and "Smallest" under "Size portfolios"), we select the best five of the fifty potential regressors and record the level of R^2. We then regress each of the correlated portfolio return series on the five regressors identified for the base portfolio and record the R^2. The simulation procedure is repeated 100 times. The ratio of the average R^2 of the correlated portfolio to the average R^2 of the base portfolio is computed and is reported below as "Expected explanatory power." If the five factors that best describe U.S. portfolio returns produce an R^2 of twenty percent, for example, the same factors are expected to produce an R^2 of eight percent for the Australian portfolio returns (i.e., the reported expected explanatory power of 0.40 times 20 percent).

Country Portfolios	Assumed Correlation	Expected Explanatory Power	Industry Portfolios	Assumed Correlation	Expected Explanatory Power	Size Portfolios	Assumed Correlation	Expected Explanatory Power
United States	1.00	1.00	Petroleum	1.00	1.00	Smallest	1.00	1.00
Australia	0.48	0.40	Finance/real estate	0.72	0.62	2	0.95	0.89
Austria	0.13	0.26	Consumer durables	0.56	0.49	3	0.92	0.85
Belgium	0.41	0.40	Basic industries	0.65	0.55	4	0.89	0.81
Canada	0.72	0.62	Foods/tobacco	0.54	0.47	5	0.87	0.78
Denmark	0.32	0.33	Construction	0.58	0.52	6	0.84	0.74
France	0.43	0.42	Capital goods	0.56	0.46	7	0.83	0.73
Germany	0.33	0.34	Transportation	0.56	0.49	8	0.79	0.68
Hong Kong	0.29	0.31	Utilities	0.57	0.46	9	0.77	0.67
Italy	0.23	0.29	Textile/trade	0.41	0.36	Largest	0.71	0.61
Japan	0.28	0.36	Services	0.54	0.48			
Holland	0.56	0.52	Leisure	0.49	0.42			
Norway	0.44	0.37						
Singapore	0.46	0.40						
Spain	0.25	0.32						
Sweden	0.39	0.38						
Switzerland	0.49	0.46						
United Kingdom	0.50	0.43						

value "expected explanatory power." In the case of the Australian portfolio, for example, the expected explanatory power ratio is 0.40, which means that, if we find a set of five factors that explains twenty percent of the variation of U.S. stock market returns, we should expect the same five factors to explain eight percent of the variation of Australian stock market returns.

The expected explanatory power ratios reported for the various country portfolios range from 0.26 for Austria to 0.62 for Canada. Not surprisingly, these were the two country portfolios with the lowest and highest assumed correlations with the U.S. portfolio. The assumed correlations for the industry portfolios are higher than the country portfolios. Consequently, the expected explanatory power ratios are higher. The ratios are highest for the size portfolios. The second smallest size portfolio, for example, has correlation of 0.95 with the smallest firm portfolio, and its expected explanatory power is 0.89. A

set of factors that explain twenty percent of the variation of small stock returns, therefore, is expected to explain about 18 percent of the variation of the returns of the second smallest size portfolio.

In summary, while out-of-sample prediction procedures may reduce the variable-selection bias, the bias cannot be fully removed because the portfolio return samples are almost surely to be correlated. And, the higher is the correlation, the less effective are the out-of-sample procedures. Consequently, the important task of determining meaningful inference when a wide array of economic time series (and their transformations) are available remains.

B. Past Studies of Return Prediction

The focus now turns to the results of past studies of monthly return predictability. We use the Bonferroni bound (3) and the Rencher/Pun rule-of-thumb (5) to provide the reader with a better understanding of how the variable-selection dilemma might affect standard significance tests. It is important to stress that we are considering only one test procedure, based on the R^2 and designed to correct the classical R^2 for the effects of exhaustive variable selection. Many other factors go into testing any predictive model. The sign and magnitude of the estimated coefficients and the restrictions across various portfolios, for example, provide important information. Nevertheless, our test of the null hypothesis that all of the β coefficients are zero is crucial.

To see whether at least one regressor in the prediction model has a coefficient different from zero, expressions (3) and (5) are used to compute 95 percent cutoff levels for m using the R^2 values reported in past research. The cutoff level, denoted m^*, may be interpreted as the minimum number of potential regressors required to achieve the reported R^2 using unrelated data. Naturally, these tests cannot tell us how variable selection was performed. They simply answer the question, how many explanatory variables would have to be examined to find an R^2 as high as the one reported five percent of the time. In studies where a model is not significant using standard classical R-squared values, the value of m^* is reported to be zero. For some models, the cutoff m^* is so high that we could not compute a numerical value. In these cases, we report the value of m^* to be ∞. To assist the reader in comparing the results of the various studies, Table IV provides a summary.[11]

Keim and Stambaugh (1986) Study

The first study that we consider is Keim and Stambaugh (1986). To predict monthly returns, they use (a) the yield on under Baa-rated bonds less the one-month T-bill yield; (b) the logarithm of the ratio given by the level of the S&P 500 index (deflated by the consumer price index) and the average of the year-end real S&P 500 index over the 45 prior years; and (c) minus the natural logarithm of share price, averaged equally across the quintile of smallest

[11] All studies except Campbell (1987) report $\bar{R}^2$ values. We find the corresponding R^2 using the relation given by Footnote 4.

Table IV

Minimum Number of Potential Regressors Required to Achieve Reported *R*-Squared Values in Past Research

This table reports the 95 percent confidence limit for m^*, the minimum number of regressors required to achieve an R-squared at least as high as that reported in a number of recent empirical studies. If the researcher had access to more than the listed m^* potential regressors, their reported R-squared would occur at least five percent of the time through random chance according to the test used. In cases where the model is not significant using standard classical R-squared values, the value of m^* is reported to be zero. In cases where we could not compute the m^* value because of constraints on numerical precision, the value of m^* is reported to be ∞. The table lists the research studies, the number of instruments used in their models, the number and type of portfolios examined, the number of observations in their data sets, their reported R-squared values, and the m^* values calculated using a Bonferroni bound and the Rencher/Pun (1980) rule-of-thumb approximation. All studies except Campbell (1987) report $\bar{R}^2$. Campbell reports R^2.

Study	No. of Instruments	No. of Portfolios	Type of Portfolio	No. of Observations	Reported R-Squared Values	95% Confidence Limit for m^*	
						Bonferroni	Rule-of-Thumb
Keim and Stambaugh (1986, pp. 369–370)	1	4	Bond	611	0.016 to 0.088	49 to ∞	50 to ∞
	1	4	Bond	300	−0.001 to 0.097	0 to ∞	0 to ∞
	1	4	Bond	311	0.007 to 0.069	0 to ∞	0 to ∞
	1	3	Stock	611	0.001 to 0.014	0 to 25	0 to 26
	1	3	Stock	300	−0.003 to 0.020	0 to 6	0 to 7
	1	3	Stock	311	−0.003 to 0.003	0	0
Campbell (1987, p. 378)	4	3	Bond	244	0.252, 0.126, 0.032	∞, 31, 0	∞, 176, 0
	4	3	Bond	51	0.231, 0.199, 0.157	0, 0, 0	0, 0, 0
	4	1	Stock	244	0.112	21	84
	4	1	Stock	51	0.228	0	0
Harvey (1989, p. 298)	5	10	Size	556	0.067 to 0.179	47 to 787	268 to 4273
	5	1	Value-weighted	556	0.075	72	462
Ferson and Harvey (1991, p. 51)							
(with January dummy)	6	12	Industry	276	0.058 to 0.137	9 to 35	11 to 146
	6	3	Size	276	0.196, 0.153, 0.105	∞, 51, 18	∞, 234, 51
	6	3	Bond	276	0.040, 0.055, 0.092	0, 9, 15	0, 10, 32
(without January dummy)	5	12	Industry	276	0.059 to 0.132	9 to 48	12 to 274
	5	3	Size	276	0.079, 0.122 0.109	13, 35, 26	28, 186,109
	5	3	Bond	276	0.038, 0.059, 0.094	6, 9, 18	6, 12, 56

market value firms on the New York Stock Exchange (NYSE). They regress each of these variables separately on four bond and three stock portfolios. Their sample consists of monthly data during the period January 1928 through November 1978. They examine the full period (611 observations), the subperiod from January 1928 through December 1952 (300 observations), and the subperiod from January 1953 through November 1978 (311 observations). They report $\bar{R}^2$ values that are scaled by the estimated first-order autocorrelation of the residuals. They also use weighted residuals in their computations.

In computing m^*, we treat the scaled $\bar{R}^2$ values as if they were the usual $\bar{R}^2$ measures. One could argue that this approach is conservative in the sense that, because the maximum autocorrelation is one, the values of $\bar{R}^2$ are inflated and the bias will be towards rejecting the null hypothesis. For the case of the residual weighting, however, the effects on the distributional assumptions underlying equation (3) are unclear.

Keim and Stambaugh use only one regressor in their models. With one regressor, the various regression specifications have neither overlapping x variables nor correlation among regressors. In such situations, expression (2) is a bound on the joint distribution and is closely approximated by the Bonferonni bound equation (3). In the full sample, the worst fitting bond equation has an $\bar{R}^2$ of 0.016 and generates a Remcher/Pun rule-of-thumb measure of m^* value equal to 50, indicating that with access to this number of regressors the reported $\bar{R}^2$ would be exceeded five percent of the time. On the other hand, the best fitting bond equation with an $\bar{R}^2$ of 0.088 yields an m^* of ∞, which means that the number of randomly chosen regressors needed to duplicate this result is so large that we cannot compute precisely how many would be needed. The stock return equations do not fit the data so well, and we find that, with $\bar{R}^2$ values ranging from 0.001 to 0.014, the associated m^* values range from zero to 26. In the subperiods, the stock return equations yield values from zero to seven. A finding that m^* is seven suggests that seven randomly chosen regressors would generate a value of $\bar{R}^2$ that exceeds the reported value of 0.228 at least five percent of the time, whereas a value of zero implies that the equation is not significant with standard cutoff R^2 values.

Campbell (1987) Study

Campbell (1987) uses (a) the one-month T-Bill rate; (b) the two-month less one-month T-Bill rate; (c) the six-month less one-month T-Bill rate; and (d) one lag of the two-month less one-month T-Bill rate as regressors. He has two nonoverlapping samples (i.e., May 1959 through August 1979 and September 1979 through November 1983) and tests his model with three bond and one stock portfolio return series.

Like in the Keim/Stambaugh study, the Campbell study shows that the best fits are obtained using the bond portfolio return series. For the first subperiod, Campbell reports R^2 values of 0.252, 0.126, and 0.032 for the three bond return series (see Table IV). The corresponding Rencher/Pun rule-of-thumb m^* values are ∞, 176, and 0. The best-fitting regression is for the two-month T-bill

portfolio returns, which produces the R^2 value of 0.252. The associated m^* value is so large that we cannot compute it accurately. The bond portfolio results in the second subperiod indicate that none of the equations are significant with standard cutoff R^2 values.

The R^2 value for the stock return regression is 0.112 in the first subperiod. The corresponding rule-of-thumb m^* value is 84. In other words, having access to 84 randomly chosen regressors would generate R^2 values of 0.112 at least five percent of the time. The reported R^2 for the second subperiod is 0.228. The corresponding m^* value is zero, considerably less than the 84 reported for the first subperiod. Among other things, this reflects smaller size—the smaller the sample size, the greater the danger of over-fitting and consequently the lower is the value of m^*.

Harvey (1989) Study

Harvey (1989) predicts the returns of eleven stock portfolios (ten size decile portfolios created from the Center for Research in Security Prices (CRSP) monthly return file as well as one value-weighted index) using (a) the excess return on equal-weighted market return; (b) the junk bond premium (Baa-Aaa bond yields); (c) the dividend-yield spread (yield on the S&P 500 stock index minus yield on a one-month T-Bill); (d) the term premium (the difference in returns for holding a 90-day bill and a 30-day bill for one month); and (e) a January dummy variable.

For the sample period September 1941 through December 1987, Harvey finds in-sample $\bar{R}^2$ values for the ten size portfolios from 0.067 to 0.179, with the $\bar{R}^2$ increasing as firm size grows small. The smallest decile portfolio has an $\bar{R}^2$ of 0.179, which means the value of m^* is 4,273. In other words, it would take an exhaustive search across 4,273 randomly chosen regressors to generate a value of $\bar{R}^2$ that exceeds the reported value of 0.179 at least five percent of the time. As firm size increases, the model fits the data less and less well. The largest stock portfolio, for example, has an $\bar{R}^2$ is 0.067, which implies a rule-of-thumb value of 268. For the value-weighted portfolio, the $\bar{R}^2$ is 0.075 and the rule-of-thumb m^* is 462.

Ferson and Harvey (1991) Study

Ferson and Harvey (1991) report $\bar{R}^2$ values for three bond portfolio and fifteen stock portfolio return prediction models fitted using monthly data for the period 1964 through 1986. As instruments, they use (a) the excess return on equally weighted market return; (b) the junk bond premium (Baa-Aaa); (c) the dividend yield (sum of the previous years dividends on the S&P 500 stock index divided by the price level in a given month); (d) the term premium (the difference in returns for holding a 90-day bill and a 30-day bill for one month); (e) the one-month nominal T-bill rate; and (f) a January dummy variable.

For the twelve industry stock portfolios, the reported $\bar{R}^2$ values range from 0.058 to 0.137. With six regressors and 276 observations, the corresponding rule-of-thumb m^* values range from eleven to 146. Without the January

dummy, the $\bar{R}^2$ values range from twelve to 274. For the three size portfolios, the reported $\bar{R}^2$ produces m^* values of ∞, 234, and 51 for the regressions including the January dummy and 28, 186, and 109 for the regression excluding the January dummy. For only one of these stock portfolio regressions is the critical number of randomly chosen regressors so high that it is impossible to duplicate the result.

For the Treasury bond, corporate bond, and Treasury bill regressions, the rule-of-thumb values are 0, 10, and 32 for the model including the January dummy. Unlike the Keim/Stambaugh (1986) and Campbell (1987) studies, the Ferson/Harvey bond portfolios do not fit the data so well. As few as 56 randomly chosen regressors will produce $\bar{R}^2$ values as high as those reported five percent of the time.

Overall, what the results summarized in Table IV indicate is that *some* security return prediction models reject the null hypothesis that all β coefficients are zero. The number of randomly chosen regressors that are required to duplicate the results is so large that an exhaustive search is implausible. On the other hand, *some* models produce m^* values that are noticeably lower. The return prediction results for the large capitalization stocks, for example, cannot reject the hypothesis that all coefficients are zero at the five percent probability level if the number of potential regressors m exceeds, say, 250. Put differently, as few as 250 randomly selected regressors together with an exhaustive search of all possible regression combinations could find regressions that would appear to do a better job of predicting security returns. Of course, this is only one method to determine whether security returns are predictable. Other, more detailed specification tests are sure to shed additional light on this issue.

IV. Summary

This article proposes alternative techniques for assessing goodness-of-fit of OLS regressions when a researcher has had access to many potential regressors (or, equivalently, has read past research that suggested which regressors to choose). Relying on the applied statistics literature, we provide two variants of the conventional F-test to determine whether at least one of the regressors used is nonzero. The Bonferroni test is conservative because it is based on the examination of a bound on the joint distribution of the R^2 statistic across $\binom{m}{k}$ regressions and will, at times, not reject the null hypothesis of all coefficients being zero when it should. The Rencher/Pun (1980) rule-of-thumb is an approximation of the distribution of the maximal R^2 and appears reasonably accurate for the numbers of parameters and observations typically used in monthly return prediction tests. Of course, with added computational cost, we can approximate the exact distribution of the maximal R^2 using Monte Carlo simulation.

The fact that the same explanatory variables may appear to work well across various country, industry, and size portfolios is not a validation of the use of the explanatory variables. Our numerical analysis shows that, when fitting a

number of portfolios simultaneously, high correlations between portfolios means high R^2 values for predictive models that use the same instruments. As a consequence, the use of other industry, size, or country data as a control to guard against variable-selection biases can be misleading.

REFERENCES

Black, F., 1992, Estimating expected returns, Working paper, Goldman, Sachs & Co.

Breiman, L., 1992, The little bootstrap and other methods for dimensionality selection in regression: X-Fixed prediction error, *Journal of the American Statistical Association* 87, 738–754.

Breiman, L., and P. Spector, 1992, Submodel selection and evaluation in regression: The X-random case, *International Statistical Review* 60, 291–319.

Campbell, J. Y., 1987, Stock returns and the term structure, *Journal of Financial Economics* 18, 373–399.

Cramer, J. S., 1987, Mean and variance of R^2 in small and moderate samples, *Journal of Econometrics* 35, 253–266.

David, H. A., 1981, *Order Statistics* (John Wiley & Sons, New York).

Denton, F., 1985, Data mining as an industry, *The Review of Economics and Statistics* 67, 124–127.

Ferson, W., and C. Harvey, 1991, Sources of predictability in portfolio returns, *Financial Analysts Journal* 47, 49–56.

Freedman, D. A., 1983, A note on screening regression equations, *The American Statistician* 37, 152–155.

Granger, C. W. J., and P. Newbold, 1974, Spurious regressions in econometrics, *Journal of Econometrics* 2, 111–120.

Harvey, C. R., 1989, Time-varying conditional covariances in tests of asset pricing models, *Journal of Financial Economics* 24, 289–317.

Hjorth, J. S. U., 1994, *Computer Intensive Statistical Methods* (Chapman & Hall, London).

Keim, D. B., and R. F. Stambaugh, 1986, Predicting returns in the stock and bond markets, *Journal of Financial Economics* 17, 357–390.

Kimball, A., 1951, On dependent tests of significance in the analysis of variance, *Annals of Mathematical Studies* 22, 600–602.

Lo, A., and A. C. MacKinlay, 1990, Data-snooping biases in tests of financial asset pricing models, *The Review of Financial Studies* 3, 431–467.

Lo, A., and A. C. MacKinlay, 1992, Maximizing predictability in the stock and bond markets, Working paper, Massachusetts Institute of Technology.

Merton, R., 1987, On the current state of the stock market rationality hypothesis, in R. Dornbusch, S. Fisher, and J. Bossons, Eds.: *Macroeconomics and Finance: Essays in Honor of Franco Modigliani* (M.I.T. Press, Cambridge).

Miller, A. J., 1984, Selection of subsets of regression variables (with discussion), *Journal of the Royal Statistical Society* UL Series A, 147, 398–425.

Miller, A. J., 1990, *Subset Selection in Regression* (Chapman & Hall, London).

Rencher, A., and F. Pun, 1980, Inflation of R^2 in best subset regressions, *Technometrics* 22, 49–53.

Ross, S. A., 1989, Regression to the max, Working paper, Yale University.