THE JOURNAL OF FINANCE • VOL. LVIII, NO. 4 • AUGUST 2003

Spurious Regressions in Financial Economics?

WAYNE E. FERSON, SERGEI SARKISSIAN, and TIMOTHY T. SIMIN*

ABSTRACT

Even though stock returns are not highly autocorrelated, there is a spurious regression bias in predictive regressions for stock returns related to the classic studies of Yule (1926) and Granger and Newbold (1974). Data mining for predictor variables interacts with spurious regression bias. The two effects reinforce each other, because more highly persistent series are more likely to be found significant in the search for predictor variables. Our simulations suggest that many of the regressions in the literature, based on individual predictor variables, may be spurious.

PREDICTIVE MODELS FOR COMMON STOCK RETURNS have long been a staple of financial economics. Early studies, reviewed by Fama (1970), used such models to examine market efficiency. Stock returns are assumed to be predictable, based on lagged instrumental variables, in the current conditional asset pricing literature. Standard lagged variables include the levels of short-term interest rates, payout-to-price ratios for stock market indexes, and yield spreads between low-grade and high-grade bonds or between long- and short-term bonds. Many of these variables behave as persistent, or highly autocorrelated, time series.

This paper studies the finite sample properties of stock return regressions with persistent lagged regressors. We focus on two issues. The first is spurious regression, analogous to Yule (1926) and Granger and Newbold (1974). These studies warned that spurious relations may be found between the levels of trending time series that are actually independent. For example, given two independent random walks, it is likely that a regression of one on the other will produce a "significant" slope coefficient, evaluated by the usual *t*-statistics.

In this paper, the dependent variables are asset rates of return, which are not highly persistent. Thus, one may think that spurious regression problems

* Ferson is at the Carroll School of Management, Boston College and is a Research Associate, National Bureau of Economic Research; Sarkissian is on the Faculty of Management, McGill University; Simin is at the Smeal College of Business, Pennsylvania State University. We are grateful to Eugene Fama for suggesting the question that motivates this research and to John Cochrane, Frank Diebold, Richard C. Green, Gordon Hanka, Raymond Kan, Donald Keim, Jeffrey Pontiff, Bill Schwert, Rossen Valkanov, and an anonymous referee for helpful comments. Ferson acknowledges financial support from the Pigott-Paccar professorship at the University of Washington and the Collins Chair in Finance at Boston College. Sarkissian acknowledges financial support from FCAR and IFM2. This paper has benefited from workshops at McGill University, at the July 2000 NBER Asset Pricing Group, the 2000 Northern Finance Association Meetings, and the 2001 American Finance Association Meetings.

are unlikely. However, the returns may be considered to be the sum of an unobserved expected return, plus unpredictable noise. If the underlying *expected* returns are persistent time series, there is still a risk of spurious regression. Because the unpredictable noise represents a substantial portion of the variance of stock returns, the spurious regression results will differ from those in the classical setting.

The second issue is "data mining" as studied for stock returns by Lo and MacKinlay (1990), Foster, Smith, and Whaley (1997), and others. If the standard instruments employed in the literature arise as the result of a collective search through the data, they may have no predictive power in the future. Stylized "facts" about the dynamic behavior of stock returns using these instruments (e.g., Cochrane (1999)) could be artifacts of the sample. Such concerns are natural, given the widespread interest in predicting stock returns.

We focus on spurious regression and the interaction between data mining and spurious regression bias. If the underlying expected return is not predictable over time, there is no spurious regression bias, even if the chosen regressor is highly autocorrelated. In this case, our analysis reduces to pure data mining as studied by Foster et al. (1997).

When expected returns are persistent, spurious regression bias calls some of the evidence of previous studies into question. We examine univariate regressions for the Standard and Poors 500 (S&P 500) excess return using 13 popular lagged instruments over the sample periods of the original studies. We find that 7 of 26 t-ratios or regression R-squares, significant by the usual five percent criteria, are consistent with the null hypothesis of a spurious regression.

The spurious regression and data mining effects reinforce each other. If researchers have mined the data for regressors that produce high "R-squares" in predictive regressions, the mining is more likely to uncover the spurious, persistent regressors. The standard regressors in the literature tend to be highly autocorrelated, as expected if the regressors result from a spurious mining process. For reasonable parameter values, all the regressions that we review from the literature are consistent with a spurious mining process, even when only a small number of instruments are considered in the mining.

This paper contributes to a substantial literature that studies the sampling properties of predictive regressions for stock returns. Goetzmann and Jorion (1993), Nelson and Kim (1993), Bekaert, Hodrick, and Marshall (1997), and Stambaugh (1999) study biases due to dependent stochastic regressors. Kim, Nelson, and Startz (1991) study structural-change-induced misspecification. Campbell and Shiller (1988) consider dependent regressors with unit roots. Fama and French (1988a), Kandel and Stambaugh (1990), and Hodrick (1992) focus on autocorrelation for long-horizon stock returns. Lanne (2001) and Valkanov (2001) develop statistical inference methods in the presence of near unit roots. Pesaran and Timmermann (1995), Bossaerts and Hillion (1999), Goyal and Welch (2002), and Simin (2002) examine model selection and out-of-sample validity. Boudoukh and Richardson (1994) provide an overview of econometric issues. Schwert (2002) reviews anomalies and trading strategies based on predictability.

The rest of the paper is organized as follows. Section I describes the data. Section II presents the models used in the simulation experiments. Section III presents the simulation results. First, we study the pure spurious regression issue in isolation. Then we consider the interaction between spurious regression and data mining biases. Section IV offers concluding remarks.

I. The Data

Table I surveys nine of the major studies that propose instruments for predicting stock returns. The table reports summary statistics for monthly data, covering various subperiods of 1926 through 1998. The sample size and period depends on the study and the variable, and the table provides the details. We attempt to replicate the data series that were used in these studies as closely as possible. The summary statistics are from our data. Note that the first-order autocorrelations frequently suggest a high degree of persistence. For example, the short-term Treasury bill yields, monthly book-to-market ratios, the dividend yield of the S&P 500, and some of the yield spreads have sample first order autocorrelations of 0.97 or higher.

Table I also summarizes regressions for the monthly return of the S&P 500 stock index, measured in excess of the one-month Treasury bill return from Ibbotson Associates, on the lagged instruments. These are OLS regressions using one instrument at a time. We report the slope coefficients, their t-ratios, and the adjusted R-squares. The R-squares range from less than one percent to more than seven percent, and 8 of the 13 t-ratios are larger than 2.0. The t-ratios are based on the OLS slopes and Newey–West (1987) standard errors, where the number of lags is chosen based on the number of statistically significant residual autocorrelations.[1]

The small R-squares in Table I suggest that predictability represents a tiny fraction of the variance in stock returns. However, even a small R-squared can signal economically significant predictability. For example, Kandel and Stambaugh (1996) and Fleming, Kirby, and Ostdiek (2001) find that optimal portfolios respond by a substantial amount to small R-squares in standard models. Studies combining several instruments in multiple regressions report higher R-squares. For example, Harvey (1989), using five instruments, reports adjusted R-squares as high as 17.9 percent for size portfolios. Ferson and Harvey (1991) report R-squares of 5.8 percent to 13.7 percent for monthly size and industry portfolio returns. These values suggest that the "true" R-squared, if we could regress the stock return on its time-varying conditional mean, might be substantially higher than we see in Table I. To accommodate this possibility, we allow the true R-squares in our simulations to vary over the range from 0 to 15 percent. For exposition we focus on an intermediate value of 10 percent.

[1] Specifically, we compute 12 sample autocorrelations and compare the values with a cutoff at two approximate standard errors: $2/\sqrt{T}$, where T is the sample size. The number of lags chosen is the minimum lag length at which no higher order autocorrelation is larger than two standard errors. The number of lags chosen is indicated in the far right column.

Table I

Common Instrumental Variables: Sources, Summary Statistics, and OLS Regression Results

This table summarizes variables used in the literature to predict stock returns. The first column indicates the published study. The second column denotes the lagged instrument. The next two columns give the sample (Period) and the number of observations (Obs) on the stock returns. Columns 5 and 6 report the autocorrelation (ρ_Z) and the standard deviation of the instrument (σ_Z), respectively. The next three columns report regression results for S&P 500 excess return on a lagged instrument. The slope coefficient is β, the t-statistic is t, and the coefficient of determination is R^2. The last column (HAC) reports the method used in computing the standard errors of the slopes. The method of Newey–West (1987) is used with the number of lags given in parentheses. MA (·) refers to the number of moving average terms used in the covariance matrix. The abbreviations in the table are as follows. *TB1y* is the yield on the one-month Treasury bill. *Two-one, Six-one,* and *Lag(two)-one* are computed as the spreads on the returns of the two- and one-month bills, six- and one-month bills, and the lagged value of the two-month and current one-month bill. The yield on all corporate bonds is denoted as *ALLy*. The yield on AAA rated corporate bonds is *AAAy*, and *UBAAy* is the yield on corporate bonds with a below BAA rating. The variable "*Cay*" is the linear function of consumption, asset wealth, and labor income. The book-to-market ratios for the Dow Jones Industrial Average and the S&P 500 are respectively *DJBM* and *SPBM*.

(1) Reference	(2) Predictor	(3) Period	(4) Obs	(5) ρ_z	(6) σ_z	(7) β	(8) t	(9) R^2	(10) HAC
Breen, Glosten, & Jagannathan (1989)	*TB1y*	5404–8612	393	0.97	0.0026	− 2.49	− 3.58	0.023	NW(5)
Campbell (1987)	*Two–one*	5906–7908	264	0.32	0.0006	11.87	2.38	0.025	NW(0)
	Six–one	5906–7908	264	0.15	0.0020	2.88	2.13	0.025	NW(0)
	Lag(two) − one	5906–7908	264	0.08	0.0010	9.88	2.67	0.063	NW(6)
Fama (1990)	*ALLy–AAAy*	5301–8712	420	0.97	0.0040	0.88	1.46	0.005	MA(0)
Fama & French (1988a)	Dividend yield	2701–8612	720	0.97	0.0013	0.40	1.36	0.007	MA(9)
Fama & French (1989)	*AAAy–TB1y*	2601–8612	732	0.92	0.0011	0.51	2.16	0.007	MA(9)
Keim & Stambaugh (1986)	*UBAAy*	2802–7812	611	0.95	0.0230	1.50	0.75	0.002	MA(9)
	UBAAy–TB1y	2802–7812	611	0.97	0.0320	1.57	1.48	0.007	MA(9)
Kothari & Shanken (1997)	*DJBM*	1927–1992	66	0.66	0.2270	0.28	2.63	0.078	MA(0)
Lettau & Ludvigson (2001)	"*Cay*"	52Q4–98Q4	184	0.79	0.0110	1.57	2.58	0.057	MA(7)
Pontiff & Schall (1998)	*DJBM*	2602–9409	824	0.97	0.2300	2.96	2.16	0.012	MA(9)
	SPBM	5104–9409	552	0.98	0.0230	9.32	1.03	0.001	MA(5)

To incorporate data mining, we compile a randomly selected sample of 500 potential instruments, through which our simulated analyst sifts to mine the data for predictor variables. We select the 500 series randomly from a much larger sample of 10,866 potential variables. The specifics are described in the Appendix. Essentially, the procedure is to generate uniformly distributed random numbers, order the series from 1 to 10,866 and randomly extract 500 series. The 500 series are randomly ordered, and permanently assigned numbers between 1 and 500. When a data miner in our simulations searches through, say 50 series, we use the sampling properties of the 50 series to calibrate the parameters in the simulations.

We also use our sample of potential instruments to calibrate the parameters that govern the amount of persistence in the "true" expected returns in the model. On the one hand, if the instruments we see in the literature, summarized in Table I, arise from a spurious mining process, they are likely to be more highly autocorrelated than the underlying "true" expected stock return. On the other hand, if the instruments in the literature are a realistic representation of expected stock returns, the autocorrelations in Table I may be a good proxy for the persistence of the true expected returns.[2] The mean autocorrelation of our 500 series is 15 percent and the median is 2 percent. Eleven of the 13 sample autocorrelations in Table I are higher than 15 percent, and the median value is 95 percent. We consider a range of values for the true autocorrelation based on these figures, as described below.

II. The Models

Consider a situation in which an analyst runs a time-series regression for the future stock return, r_{t+1}, on a lagged predictor variable:

$$r_{t+1} = \alpha + \delta Z_t + v_{t+1}. \tag{1}$$

The data are actually generated by an unobserved latent variable, Z_t^*, as

$$r_{t+1} = \mu + Z_t^* + u_{t+1}, \tag{2}$$

where u_{t+1} is white noise with variance, σ_u^2. We interpret the latent variable, Z_t^* as the deviations of the conditional mean return from the unconditional mean, μ, where the expectations are conditioned on an unobserved "market" information set at time t. The predictor variables follow an autoregressive process:

$$(Z_t^*, Z_t)' = \begin{Bmatrix} \rho^* & 0 \\ 0 & \rho \end{Bmatrix} (Z_{t-1}^*, Z_{t-1})' + (\varepsilon_t^*, \varepsilon_t)' \tag{3}$$

[2] There are good reasons to think that expected stock returns may be persistent. Asset pricing models like the consumption model of Lucas (1978) describe expected stock returns as functions of expected economic growth rates. Merton (1973) and Cox, Ingersoll, and Ross (1985) propose real interest rates as candidate state variables, driving expected returns in intertemporal models. Such variables are likely to be highly persistent. Empirical models for stock return dynamics frequently involve persistent, autoregressive expected returns (e.g., Conrad and Kaul (1988), Fama and French (1988b), Lo and MacKinlay (1988), or Huberman and Kandel (1990)).

The assumption that the true expected return is autoregressive follows previous studies such as Conrad and Kaul (1988), Fama and French (1988b), Lo and MacKinlay (1988), and Huberman and Kandel (1990).

To generate the artificial data, the errors $(\varepsilon_t^*, \varepsilon_t)$ are drawn randomly as a normal vector with mean zero and covariance matrix, Σ. We build up the time-series of the Z and Z^* through the vector autoregression equation (3), where the initial values are drawn from a normal with mean zero and variances, Var(Z) and Var(Z^*). The other parameters that calibrate the simulations are $\{\mu, \sigma_u^2, \rho, \rho^*, \text{and } \Sigma\}$.

We have a situation in which the "true" returns may be predictable, if Z_t^* could be observed. This is captured by the *true* R-squared, $\text{Var}(Z^*)/[\text{Var}(Z^*) + \sigma_u^2]$. We set Var($Z^*$) to equal the sample variance of the S&P 500 return, in excess of a one-month Treasury bill return, multiplied by 0.10. When the true R-squared of the simulation is 10 percent, the unconditional variance of the r_{t+1} that we generate is equal to the sample variance of the S&P 500 return, and the first-order autocorrelation is similar to that of the actual data. When we choose other values for the true R-squared, these determine the values for the parameter σ_u^2. We set μ to equal the sample mean excess return of the S&P 500 over the 1926 through 1998 period, or 0.71 percent per month.

The extent of the spurious regression bias depends on the parameters ρ and ρ^*, which control the persistence of the measured and the true regressor. These values are determined by reference to Table I and from our sample of 500 potential instruments. The specifics differ across the special cases, as described below.

While the stock return could be predicted if Z_t^* could be observed, the analyst uses the measured instrument Z_t. If the covariance matrix Σ is diagonal, Z_t and Z_t^* are independent, and the true value of δ in the regression (1) is zero.

A. Pure Spurious Regression

To focus on spurious regression in isolation, we specialize equation (3) as follows. The covariance matrix Σ is a 2×2 diagonal matrix with variances (σ_*^2, σ^2). For a given value of ρ^* the value of σ_*^2 is determined as $\sigma_*^2 = (1 - \rho_*^2)\text{Var}(Z^*)$. The measured regressor has $\text{Var}(Z) = \text{Var}(Z^*)$. The autocorrelation parameters, $\rho^* = \rho$ are allowed to vary over a range of values. (We also allow ρ and ρ^* to differ from one another, as described below.)

Following Granger and Newbold (1974), we interpret a spurious regression as one in which the "t-ratios" in regression (1) are likely to indicate a significant relation when the variables are really independent. The problem may come from the numerator or the denominator of the t-ratio: The coefficient or its standard error may be biased. As in Granger and Newbold, the problem lies with the standard errors.[3] The reason is simple to understand. When the null

[3] While Granger and Newbold (1974) do not study the slopes and standard errors to identify the separate effects, our simulations, designed to mimic their setting (not reported in the tables), confirm that their slopes are well behaved, while the standard errors are biased. Granger and Newbold use OLS standard errors, while we focus on the heteroskedasticity and autocorrelation-consistent standard errors that are more common in recent studies.

hypothesis that the regression slope $\delta = 0$ is true, the error term u_{t+1} of regression Equation (1) inherits autocorrelation from the dependent variable. Assuming stationarity, the slope coefficient is consistent, but standard errors that do not account for the serial dependence correctly are biased.

Because the spurious regression problem is driven by biased estimates of the standard error, the choice of standard error estimator is crucial. In our simulation exercises, it is possible to find an efficient unbiased estimator, since we know the "true" model that describes the regression error. Of course, this will not be known in practice. To mimic the practical reality, the analyst in our simulations uses the popular autocorrelation-heteroskedasticity-consistent (HAC) standard errors from Newey and West (1987), with an automatic lag selection procedure. The number of lags is chosen by computing the autocorrelations of the estimated residuals and truncating the lag length when the sample autocorrelations become "insignificant" at longer lags. (The exact procedure is described in Footnote 1, and modifications to this procedure are discussed below.)

This setting is related to Phillips (1986) and Stambaugh (1999). Phillips derives asymptotic distributions for the OLS estimators of the regression (1), in the case where $\rho = 1$, $u_{t+1} \equiv 0$, and $\{\varepsilon_t^*, \varepsilon_t\}$ are general independent mean zero processes. We allow a nonzero variance of u_{t+1} to accommodate the large noise component of stock returns. We assume $\rho < 1$ to focus on stationary, but possibly highly autocorrelated, regressors.

Stambaugh (1999) studies a case where the errors $\{\varepsilon_t^*, \varepsilon_t\}$ are perfectly correlated, or equivalently, the analyst observes and uses the correct lagged stochastic regressor. A bias arises when the correlation between u_{t+1} and ε_{t+1}^* is not zero, related to the well-known small sample bias of the autocorrelation coefficient (e.g., Kendall (1954)). In the pure spurious regression case studied here, the observed regressor Z_t is independent of the true regressor $Z_{\pm}^*$, and u_{t+1} is independent of ε_{t+1}^*. The Stambaugh bias is zero in this case. The point is that there remains a problem in predictive regressions, in the absence of the bias studied by Stambaugh, because of spurious regression.

B. Spurious Regression and Data Mining

We consider the interaction between spurious regression and data mining, where the instruments to be mined are independent as in Foster et al. (1997). There are L measured instruments over which the analyst searches for the "best" predictor, based on the R-squares of univariate regressions. In Equation (3) Z_t becomes a vector of length L, where L is the number of instruments through which the analyst sifts. The error terms $(\varepsilon_t^*, \varepsilon_t)$ become an $L+1$ vector with a diagonal covariance matrix; thus, ε_t^* is independent of ε_t.

The persistence parameters in Equation (3) become an $(L+1)$-square, diagonal matrix, with the autocorrelation of the true predictor equal to ρ^*. The value of ρ^* is either the average from our sample of 500 potential instruments, 15 percent, or the median value from Table I, 95 percent. The remaining autocorrelations, denoted by the L-vector ρ, are set equal to the autocorrelations of the first L

instruments in our sample of 500 potential instruments, when $\rho^* = 15$ percent.[4] When $\rho^* = 95$ percent, we rescale the autocorrelations to center the distribution at 0.95 while preserving the range in the original data.[5] The simulations match the unconditional variances of the instruments, Var(Z), to the data. The first element of the covariance matrix Σ is equal to σ_*^2. For a typical ith diagonal element of Σ, denoted by σ_i, the elements of $\rho(Z_i)$ and $\mathrm{Var}(Z_i)$ are given by the data, and we set $\sigma_i^2 = [1 - \rho(Z_i)^2]\mathrm{Var}(Z_i)$.

III. Simulation Results

We first consider spurious regression in isolation. Then we study spurious regression with data mining.

A. *Pure Spurious Regression*

Table II summarizes the results for the case of pure spurious regression. We record the estimated slope coefficient in regression (1), its Newey–West t-ratio, and the coefficient of determination at each trial and summarize their empirical distributions. The experiments are run for two sample sizes, based on the extremes in Table I. These are $T = 66$ and $T = 824$ in Panels A and B, respectively. In Panel C, we match the sample sizes to the studies in Table I. In each case, 10,000 trials of the simulation are run; 50,000 trials produces similar results.

The rows of Table II refer to different values for the true R-squares. The smallest value is 0.1 percent, where the stock return is essentially unpredictable, and the largest value is 15 percent. The columns of Table II correspond to different values of ρ^*, the autocorrelation of the true expected return, which runs from 0.00 to 0.99. In these experiments, we set $\rho = \rho^*$. The subpanels labeled Critical t-statistic and Critical estimated R^2 report empirical critical values from the 10,000 simulated trials, so that 2.5 percent of the t-statistics or five percent of the R-squares lie above these values.

The subpanels labeled Mean δ report the average slope coefficients over the 10,000 trials. The mean estimated values are always small, and very close to the true value of zero at the larger sample size. This confirms that the slope coefficient estimators are well behaved, so that bias due to spurious regression comes from the standard errors.

[4] We calibrate the true autocorrelations in the simulations to the sample autocorrelations, adjusted for first-order finite-sample bias as: $\hat{\rho} + (1 + 3\hat{\rho})/T$, where $\hat{\rho}$ is the OLS estimate of the autocorrelation and T is the sample size.

[5] The transformation is as follows. In the 500 instruments, the minimum bias-adjusted autocorrelation is -0.571, the maximum is 0.999, and the median is 0.02. We center the transformed distribution about the median in Table I, which is 0.95. If the original autocorrelation ρ is less than the median, we transform it to

$$.95 + (\rho - 0.02)\{(0.95 + 0.571)/(0.02 + 0.571)\}.$$

If the value is above the median, we transform it to

$$.95 + (\rho - 0.02)\{(0.999 - 0.95)/(0.999 - 0.02)\}.$$

When $\rho^* = 0$, and there is no persistence in the true expected return, the spurious regression phenomenon is not a concern. This is true even when the measured regressor is highly persistent. (We confirm this with additional simulations, not reported in the tables, where we set $\rho^* = 0$ and vary ρ.) The logic is that when the slope in Equation (1) is zero and $\rho^* = 0$, the regression error has no persistence, so the standard errors are well behaved. This implies that spurious regression is not a problem from the perspective of testing the null hypothesis that expected stock returns are unpredictable, even if a highly autocorrelated regressor is used.

Table II shows that spurious regression bias does not arise to any serious degree, provided ρ^* is 0.90 or less, and the true R^2 is one percent or less. For these parameters, the empirical critical values for the t-ratios are 2.48 ($T = 66$, Panel A), and 2.07 ($T = 824$, Panel B). The empirical critical R-squares are close to their theoretical values. For example, for a five percent test with $T = 66$ (824) the F distribution implies critical R-squared values of 5.9 percent (0.5 percent). The values in Table II when $\rho^* = 0.90$ and true $R^2 = 1$ percent, are 6.2 percent (0.5 percent); thus, the empirical distributions do not depart far from the standard rules of thumb.

Variables like short-term interest rates and dividend yields typically have first-order sample autocorrelations in excess of 0.95, as we saw in Table I. We find substantial biases when the regressors are highly persistent. Consider the plausible scenario with a sample of $T = 824$ observations where $\rho = 0.98$ and true $R^2 = 10$ percent. In view of the spurious regression phenomenon, an analyst who was not sure that the correct instrument is being used and who wanted to conduct a 5 percent, two-tailed t-test for the significance of the measured instrument would have to use a t-ratio of 3.6. The coefficient of determination would have to exceed 2.2 percent to be significant at the 5 percent level. These cutoffs are substantially more stringent than the usual rules of thumb.

Panel C of Table II revisits the evidence from the literature in Table I. The critical values for the t-ratios and R-squares are reported, along with the theoretical critical values for the R-squares implied by the F distribution. We set the true R-squared value equal to 10 percent and $\rho^* = \rho$ in each case. We find that 7 of the 17 statistics in Table I that would be considered significant using the traditional standards, are no longer significant in view of the spurious regression bias.

While Panels A and B of Table II show that spurious regression can be a problem in stock return regressions, Panel C finds that accounting for spurious regression changes the inferences about specific regressors that were found to be significant in previous studies. In particular, we question the significance of the term spread in Fama and French (1989), on the basis of either the t-ratio or the R-squared of the regression. Similarly, the book-to-market ratio of the Dow Jones index, studied by Pontiff and Schall (1998) fails to be significant with either statistic. Several other variables are marginal, failing on the basis of one but not both statistics. These include the short-term interest rate (Fama and Schwert (1977), using the more recent sample of Breen, Glosten, and Jagannathan (1989)), the dividend yield (Fama and French (1988a)), and the quality-related

yield spread (Keim and Stambaugh (1986)). All of these regressors would be considered significant using the standard cutoffs.

It is interesting to note that the biases documented in Table II do not always diminish with larger sample sizes; in fact, the critical *t*-ratios are larger in the lower right corner of the panels when $T = 824$ than when $T = 66$. The mean values of the slope coefficients are closer to zero at the larger sample size, so the larger critical values are driven by the standard errors. A sample as large as $T = 824$ is not by itself a cure for the spurious regression bias. This is typical of spurious regression with a unit root, as discussed by Phillips (1986) for infinite sample sizes and nonstationary data.[6] It is interesting to observe similar patterns, even with stationary data and finite samples.

Phillips (1986) shows that the sample autocorrelation in the regression studied by Granger and Newbold (1974) converges in limit to 1.0. However, we find only mildly inflated residual autocorrelations (not reported in the tables) for stock return samples as large as $T = 2000$, even when we assume values of the true R^2 as large as 40 percent. Even in these extreme cases, none of the empirical critical values for the residual autocorrelations are larger than 0.5. Since $u_{t+1} = 0$ in the cases studied by Phillips, we expect to see explosive autocorrelations only when the true R^2 is very large. When R^2 is small, the white noise component of the returns serves to dampen the residual autocorrelation. Thus, we are not likely to see large residual autocorrelations in asset pricing models, even where spurious regression is a problem. The residuals-based diagnostics for spurious regression, such as the Durbin–Watson tests suggested by Granger and Newbold, are not likely to be very powerful in asset pricing regressions. For the same reason, naive application of the Newey–West procedure, where the number of lags is selected by examining the residual autocorrelations, is not likely to resolve the spurious regression problem.

Newey and West (1987) show that their procedure is consistent when the number of lags used grows without bound as the sample size T increases, provided that the number of lags grows no faster than $T^{1/4}$. The lag selection procedure in Table II examines 12 lags. Even though no more than nine lags are selected for the actual data in Table I, more lags would sometimes be selected in the simulations, and an inconsistency results from truncating the lag length.[7] However, in finite samples, an increase in the number of lags can make things worse. When "too many" lags are used, the standard error estimates become excessively noisy, which thickens the tails of the sampling distribution of the *t*-ratios. This occurs

[6] Phillips derives asymptotic distributions for the OLS estimators of equation (1), in the case where $\rho = 1$, $u_{t+1} \equiv 0$. He shows that the *t*-ratio for δ diverges for large T, while $t(\delta)/\sqrt{T}$, δ, and the coefficient of determination converge to well-defined random variables. Marmol (1998) extends these results to multiple regressions with partially integrated processes, and provides references to more recent theoretical literature. Phillips (1998) reviews analytical tools for asymptotic analysis when nonstationary series are involved.

[7] At very large sample sizes, a huge number of lags can control the bias. We verify this by examining samples as large as $T = 5000$, letting the number of lags grow to 240. With 240 lags, the critical *t*-ratio when the true $R^2 = 10$ percent and $\rho = 0.98$ falls from 3.6 in Panel B of Table II to a reasonably well-behaved value of 2.23.

Table II

The Monte Carlo Simulation Results for Regressions with a Lagged Predictor Variable

The table reports the 97.5 percentile of the Monte Carlo distribution of 10,000 Newey–West t-statistics, the 95 percentile for the estimated coefficients of determination, and the average estimated slopes from the regression

$$r_{t+1} = \alpha + \delta Z_t + v_{t+1},$$

where r_{t+1} is the excess return, Z_t is the predictor variable, and $t = 1, \ldots, T$. The parameter ρ^* is the autocorrelation coefficient of the predictors, Z_t^* and Z_t. The R^2 is the coefficient of determination from the regression of excess returns r_{t+1} on the unobserved, true instrument Z_t^*. Panel A depicts the results for $T = 66$ and Panel B for $T = 824$. Panel C gives the simulation results for the number of observations and the autocorrelations in Table I. In Panel C, the true R^2 is set to 0.1. The theoretical critical R^2 is from the F-distribution.

Panel A: 66 Observations						
R^2/ρ^*	0	0.5	0.9	0.95	0.98	0.99
			Mean δ			
0.001	− 0.0480	− 0.0554	− 0.0154	− 0.0179	− 0.0312	− 0.0463
0.005	− 0.0207	− 0.0246	− 0.0074	− 0.0088	− 0.0137	− 0.0193
0.010	− 0.0142	− 0.0173	− 0.0055	− 0.0066	− 0.0096	− 0.0129
0.050	− 0.0055	− 0.0075	− 0.0029	− 0.0037	− 0.0040	− 0.0042
0.100	− 0.0033	− 0.0051	− 0.0023	− 0.0030	− 0.0026	− 0.0021
0.150	− 0.0024	− 0.0040	− 0.0020	− 0.0026	− 0.0020	− 0.0012
			Critical t-statistic			
0.001	2.1951	2.3073	2.4502	2.4879	2.4746	2.4630
0.005	2.2033	2.3076	2.4532	2.5007	2.5302	2.5003
0.010	2.2121	2.3123	2.4828	2.5369	2.5460	2.5214
0.050	2.2609	2.3335	2.6403	2.7113	2.7116	2.6359
0.100	2.2847	2.3702	2.8408	2.9329	2.9043	2.7843
0.150	2.2750	2.3959	3.0046	3.1232	3.0930	2.9417
			Critical estimated R^2			
0.001	0.0593	0.0575	0.0598	0.0599	0.0610	0.0600
0.005	0.0590	0.0578	0.0608	0.0607	0.0616	0.0604
0.010	0.0590	0.0579	0.0619	0.0623	0.0630	0.0612
0.050	0.0593	0.0593	0.0715	0.0737	0.0703	0.0673
0.100	0.0600	0.0622	0.0847	0.0882	0.0823	0.0766
0.150	0.0600	0.0649	0.0994	0.1032	0.0942	0.0850
Panel B: 824 Observations						
			Mean δ			
0.001	0.0150	0.0106	0.0141	0.0115	0.0053	− 0.0007
0.005	0.0067	0.0049	0.0069	0.0055	0.0021	− 0.0011
0.010	0.0048	0.0035	0.0052	0.0040	0.0014	− 0.0012
0.050	0.0021	0.0017	0.0029	0.0021	0.0003	− 0.0014
0.100	0.0015	0.0013	0.0023	0.0016	0.0001	− 0.0014
0.150	0.0012	0.0011	0.0021	0.0014	− 0.0000	− 0.0014

Table II— *Continued*

Panel B: 824 Observations

$R^2/\rho*$	0	0.5	0.9	0.95	0.98	0.99
			Critical t-statistic			
0.001	1.9861	2.0263	2.0362	2.0454	2.0587	2.0585
0.005	1.9835	2.0297	2.0429	2.1123	2.1975	2.2558
0.010	1.9759	2.0279	2.0655	2.1479	2.3578	2.4957
0.050	1.9878	2.0088	2.2587	2.5685	3.1720	3.7095
0.100	1.9862	2.0320	2.3758	2.7342	3.6356	4.4528
0.150	2.0005	2.0246	2.4164	2.8555	3.8735	4.9151
			Critical estimated R^2			
0.001	0.0046	0.0047	0.0047	0.0047	0.0049	0.0049
0.005	0.0046	0.0047	0.0048	0.0051	0.0056	0.0059
0.010	0.0046	0.0047	0.0050	0.0054	0.0065	0.0073
0.050	0.0046	0.0047	0.0066	0.0085	0.0132	0.0183
0.100	0.0047	0.0049	0.0084	0.0125	0.0220	0.0316
0.150	0.0046	0.0050	0.0104	0.0166	0.0308	0.0450

Panel C: Table I simulation

Obs	$\rho*$	Critical Theoretical R^2	Critical t-statistic	Critical Estimated R^2
393	0.97	0.0098	3.2521	0.0311
264	0.32	0.0146	2.0645	0.0151
264	0.15	0.0146	2.0560	0.0151
264	0.08	0.0146	2.0318	0.0146
420	0.97	0.0092	3.2734	0.0304
720	0.97	0.0053	3.2005	0.0194
732	0.92	0.0053	2.3947	0.0103
611	0.95	0.0063	2.8843	0.0167
611	0.97	0.0063	3.2488	0.0219
66	0.66	0.0586	2.4221	0.0656
184	0.79	0.0209	2.2724	0.0270
824	0.97	0.0047	3.1612	0.0173
552	0.98	0.0070	3.6771	0.0293

for the experiments in Table II. For example, letting the procedure examine 36 autocorrelations to determine the lag length (the largest number we find mentioned in published studies), the critical t-ratio in Panel A, for true $R^2 = 10$ percent and $\rho* = 0.98$, increases from 2.9 to 4.8. Nine of the 17 statistics from Table I that are significant by the usual rules of thumb now become insignificant. The results calling these studies into question are even stronger than before. Thus, simply increasing the number of lags in the Newey–West procedure does not resolve the finite sample, spurious regression bias.[8]

[8] We conduct several experiments letting the number of lags examined be 24, 36, or 48, when $T = 66$ and $T = 824$. When $T = 66$, the critical t-ratios are always larger than the values in Table II. When $T = 824$, the effects are small and of mixed sign. The most extreme reduction in a critical t-ratio, relative to Table II, is with 48 lags, true $R^2 = 15$ percent, and $\rho* = 0.99$, where the critical value falls from 4.92 to 4.23.

We draw several conclusions about spurious regression in stock return regressions. Given persistent expected returns, spurious regression can be a serious concern well outside the classic setting of Yule (1926) and Granger and Newbold (1974). Stock returns, as the dependent variable, are much less persistent than the levels of most economic time series. Yet, when the *expected* returns are persistent, there is a risk of spurious regression bias. The regression residuals may not be highly autocorrelated, even when spurious regression bias is severe. Given inconsistent standard errors, spurious regression bias is not avoided with large samples. Accounting for spurious regression bias, we find that 7 of the 17 t-statistics and regression R-squares from previous studies that would be significant by standard criteria are no longer significant.

B. Spurious Regression and Data Mining

We now consider the interaction between spurious regression and data mining. Table III summarizes the results. The columns of Panels A through D correspond to different numbers of potential instruments, through which the analyst sifts to find the regression that delivers the highest sample R-squared. The rows refer to the different values of the true R-squared.

The cases with true $R^2 = 0$ refer to data mining only, similar to Foster et al. (1997). The columns where $L = 1$ correspond to pure spurious regression bias. We hold fixed the persistence parameter for the true expected return, ρ^*, while allowing ρ to vary depending on the measured instrument. When $L = 1$, we set $\rho = 15$ percent. We consider two values for ρ^*, 15 percent or 95 percent.

Panels A and B of Table III show that when $L = 1$ and $\rho^* = 15$ percent, there is no data mining, and, consistent with Table II, there is no spurious regression problem. The empirical critical values for the t-ratios and R-squared statistics are close to their theoretical values under normality. For larger values of L and $\rho^* = 15$ percent, there is data mining, and the critical values are close to the values reported by Foster et al. (1997) for similar sample sizes.[9] There is little difference in the results for the various true R-squares. Thus, with little persistence, there is no spurious regression problem, and no interaction with data mining.

Panels C and D of Table III tell a different story. When the underlying expected return is persistent ($\rho^* = 0.95$), there is a spurious regression bias. When $L = 1$, we have spurious regression only. The critical t-ratio in Panel C increases from 2.3 to 2.8 as the true R-squared goes from 0 to 15 percent. The bias is less pronounced here than in Table II, with $\rho = \rho^* = 0.95$, which illustrates that for a given value of ρ^*, spurious regression is worse for larger values of ρ.

Spurious regression bias interacts with data mining. Consider the extreme corners of Panel C. Whereas with $L = 1$, the critical t-ratio increases from 2.3 to 2.8 as the true R-squared goes from 0 to 15 percent, with $L = 250$, the critical t-ratio increases from 5.2 to 6.3 as the true R-squared is increased. Thus, data mining magnifies the effects of the spurious regression bias. When more instruments

[9] Our sample sizes, T, are not the same as in Foster et al. (1997). When we run the experiments for their sample sizes, we closely approximate the critical values that they report.

Table III

The Monte Carlo Simulation Results of Regressions with Spurious Regression and Data Mining, with Independent Regressors

The table reports the 97.5 percentile of the Monte Carlo distribution of 10,000 Newey–West t-statistics, the 95 percentile for the estimated coefficients of determination, and the average estimated slopes from the regression

$$r_{t+1} = \alpha + \delta Z_t + v_{t+1},$$

where r_{t+1} is the excess return, Z_t is the predictor variable, and $t = 1,\ldots,T$. The R^2 is the coefficient of determination from the regression of excess returns r_{t+1} on the unobserved, true instrument Z_t^*, which has the autocorrelation ρ^*. The parameter L is the number of instruments mined, where the one with the highest estimated R^2 is chosen. Panels A and B depict the results for $T = 66$ and $T = 824$, respectively, when the autocorrelation of the true predictor, $\rho^* = 0.15$. Panels C and D depict the results for $T = 66$ and $T = 824$, respectively, when the autocorrelation of the true predictor, $\rho^* = 0.95$, the median autocorrelation in Table I. In Panel E, the true R^2 is set to 0.1 and the original distribution of instruments is transformed so that their median autocorrelation is set at 0.95. The left-hand side of Panel E gives the critical L for the given number of observations and autocorrelation that is sufficient to generate critical t-statistics or R^2s in excess of the corresponding statistics in Table I. The right-hand side of Panel E gives the critical L that is sufficient to generate critical t-statistics or R^2s in excess of the corresponding statistics in Table I when $\rho^* = 0.95$.

Panel A: 66 Observations; $\rho^* = 0.15$

R^2/L	1	5	10	25	50	100	250
				Mean δ			
0	− 0.0004	0.0002	− 0.0002	0.0004	− 0.0001	0.0001	0.0005
0.001	− 0.0114	0.0044	− 0.0069	0.0208	− 0.0078	0.0012	0.0162
0.005	− 0.0050	0.0017	− 0.0017	0.0113	− 0.0014	− 0.0031	0.0109
0.010	− 0.0035	0.0008	− 0.0014	0.0076	− 0.0002	− 0.0011	0.0098
0.050	− 0.0014	0.0004	− 0.0004	0.0018	− 0.0023	− 0.0013	0.0063
0.100	− 0.0009	0.0006	− 0.0004	0.0014	− 0.0013	− 0.0007	0.0044
0.150	− 0.0007	0.0007	− 0.0002	0.0009	− 0.0010	− 0.0010	0.0035
				Critical t-statistic			
0	2.2971	3.2213	3.5704	4.1093	4.4377	4.8329	5.2846
0.001	2.2819	3.2105	3.5418	4.1116	4.4351	4.8238	5.2803
0.005	2.2996	3.2250	3.5466	4.1190	4.4604	4.7951	5.2894
0.010	2.2981	3.2109	3.5492	4.1198	4.4728	4.7899	5.2900
0.050	2.2950	3.2416	3.5096	4.0981	4.4036	4.8803	5.2527
0.100	2.3175	3.2105	3.5316	4.1076	4.4563	4.8772	5.2272
0.150	2.3040	3.2187	3.5496	4.0644	4.5090	4.8984	5.2948
				Critical estimated R^2			
0	0.0594	0.0974	0.1153	0.1387	0.1548	0.1738	0.1944
0.001	0.0589	0.0969	0.1149	0.1386	0.1546	0.1739	0.1944
0.005	0.0591	0.0972	0.1151	0.1383	0.1545	0.1734	0.1948
0.010	0.0592	0.0967	0.1158	0.1386	0.1544	0.1733	0.1950
0.050	0.0596	0.0970	0.1163	0.1390	0.1557	0.1738	0.1955
0.100	0.0608	0.0969	0.1165	0.1392	0.1570	0.1738	0.1954
0.150	0.0612	0.0975	0.1165	0.1397	0.1577	0.1745	0.1967

Table III— *Continued*

R^2/L	1	5	10	25	50	100	250
Panel B: 824 Observations; $\rho^* = 0.15$							
Mean δ							
0	0.0000	0.0000	0.0000	0.0000	− 0.0001	− 0.0002	0.0000
0.001	− 0.0004	0.0032	− 0.0017	0.0000	− 0.0028	− 0.0058	0.0015
0.005	− 0.0002	0.0012	− 0.0004	0.0000	− 0.0020	− 0.0031	0.0007
0.010	− 0.0001	0.0009	− 0.0004	− 0.0003	− 0.0015	− 0.0020	0.0004
0.050	− 0.0001	0.0005	0.0000	− 0.0005	− 0.0006	− 0.0009	0.0004
0.100	0.0000	0.0005	− 0.0001	− 0.0003	− 0.0001	− 0.0002	0.0003
0.150	0.0000	0.0003	− 0.0003	− 0.0003	0.0001	− 0.0002	0.0002
Critical t-statistic							
0	2.0283	2.5861	2.8525	3.1740	3.3503	3.5439	3.8045
0.001	2.0369	2.6000	2.8534	3.1785	3.3616	3.5443	3.7928
0.005	2.0334	2.6043	2.8565	3.1769	3.3625	3.5440	3.7906
0.010	2.0310	2.6152	2.8694	3.1782	3.3544	3.5477	3.7917
0.050	2.0272	2.6229	2.8627	3.1846	3.3450	3.5552	3.8039
0.100	2.0115	2.6304	2.8705	3.1807	3.3648	3.5673	3.8041
0.150	2.0044	2.6327	2.8618	3.1766	3.3691	3.5723	3.7965
Critical estimated R^2							
0	0.0047	0.0079	0.0096	0.0116	0.0130	0.0145	0.0166
0.001	0.0047	0.0079	0.0096	0.0116	0.0130	0.0145	0.0166
0.005	0.0047	0.0080	0.0096	0.0116	0.0129	0.0145	0.0166
0.010	0.0047	0.0080	0.0096	0.0115	0.0129	0.0145	0.0166
0.050	0.0047	0.0081	0.0096	0.0116	0.0130	0.0145	0.0167
0.100	0.0047	0.0081	0.0097	0.0117	0.0131	0.0146	0.0168
0.150	0.0047	0.0082	0.0096	0.0117	0.0130	0.0146	0.0168
Panel C: 66 Observations; $\rho^* = 0.95$							
Mean δ							
0	− 0.0005	0.0002	0.0006	− 0.0001	− 0.0006	− 0.0003	0.0017
0.001	− 0.0140	0.0069	0.0212	− 0.0105	− 0.0134	− 0.0112	0.0557
0.005	− 0.0060	0.0042	0.0082	− 0.0068	− 0.0024	− 0.0033	0.0240
0.010	− 0.0042	0.0031	0.0051	− 0.0029	− 0.0018	− 0.0027	0.0145
0.050	− 0.0016	0.0006	0.0035	− 0.0023	− 0.0016	− 0.0019	0.0012
0.100	− 0.0010	− 0.0002	0.0021	− 0.0013	− 0.0017	− 0.0005	0.0028
0.150	− 0.0007	− 0.0005	0.0015	− 0.0008	− 0.0011	− 0.0001	0.0013
Critical t-statistic							
0	2.3446	3.3507	3.6827	4.1903	4.4660	4.9412	5.2493
0.001	2.3641	3.3547	3.6776	4.1756	4.5157	4.9201	5.2441
0.005	2.4030	3.3864	3.7013	4.1984	4.5625	4.9381	5.2760
0.010	2.3939	3.4197	3.7308	4.1952	4.6039	4.9718	5.3083
0.050	2.5486	3.5482	3.9676	4.4703	4.9512	5.2027	5.5539
0.100	2.6955	3.7336	4.1899	4.7485	5.2335	5.5027	5.9006
0.150	2.8484	3.9724	4.4329	4.9748	5.5547	5.8256	6.2563

Table III— *Continued*

Panel C: 66 Observations; $\rho^* = 0.95$							
R^2/L	1	5	10	25	50	100	250
Critical estimated R^2							
0	0.0579	0.0974	0.1140	0.1374	0.1515	0.1689	0.1885
0.001	0.0587	0.0981	0.1143	0.1376	0.1518	0.1692	0.1884
0.005	0.0596	0.0987	0.1153	0.1385	0.1530	0.1699	0.1895
0.010	0.0604	0.1002	0.1166	0.1402	0.1543	0.1711	0.1910
0.050	0.0691	0.1113	0.1307	0.1552	0.1711	0.1859	0.2057
0.100	0.0802	0.1265	0.1508	0.1774	0.1952	0.2099	0.2307
0.150	0.0911	0.1451	0.1728	0.2021	0.2209	0.2370	0.2587
Panel D: 824 Observations; $\rho^* = 0.95$							
Mean δ							
0	− 0.0001	0.0000	0.0000	0.0000	0.0001	0.0002	0.0001
0.001	− 0.0027	− 0.0016	− 0.0007	0.0005	0.0015	0.0072	0.0039
0.005	− 0.0012	− 0.0004	0.0003	0.0006	− 0.0008	0.0029	0.0026
0.010	− 0.0009	− 0.0005	0.0000	0.0003	− 0.0008	0.0013	0.0006
0.050	− 0.0004	− 0.0005	0.0001	− 0.0002	0.0007	− 0.0006	0.0001
0.100	− 0.0003	− 0.0002	− 0.0001	− 0.0003	0.0000	0.0001	− 0.0004
0.150	− 0.0003	0.0000	0.0000	− 0.0002	0.0001	0.0002	− 0.0002
Critical t-statistic							
0	1.9807	2.6807	2.8535	3.1579	3.3640	3.5673	3.8103
0.001	1.9989	2.6876	2.8758	3.1745	3.3702	3.5792	3.8252
0.005	2.0406	2.7588	2.9269	3.2218	3.4497	3.6493	3.9075
0.010	2.1108	2.8538	3.0150	3.3500	3.5548	3.7836	4.0351
0.050	2.4338	3.3118	3.6292	4.1202	4.3685	4.6795	4.9741
0.100	2.6274	3.6661	4.0003	4.5660	4.9129	5.2567	5.6937
0.150	2.7413	3.8720	4.2048	4.8481	5.2200	5.5846	6.0420
Critical estimated R^2							
0	0.0045	0.0080	0.0096	0.0113	0.0129	0.0145	0.0164
0.001	0.0046	0.0082	0.0097	0.0115	0.0130	0.0146	0.0167
0.005	0.0048	0.0086	0.0102	0.0121	0.0137	0.0153	0.0176
0.010	0.0050	0.0092	0.0108	0.0131	0.0146	0.0163	0.0187
0.050	0.0077	0.0145	0.0173	0.0216	0.0244	0.0273	0.0314
0.100	0.0113	0.0216	0.0264	0.0331	0.0374	0.0421	0.0482
0.150	0.0151	0.0293	0.0356	0.0446	0.0508	0.0568	0.0647

Panel E: Table I Simulation

Obs	ρ^*	Critical L (t-statistic)	Critical L (R^2)	ρ^*	Critical L (t-statistic)	Critical L (R^2)
393	0.97	2	1	0.95	4	2
264	0.32	2	5	0.95	1	1
264	0.15	2	5	0.95	1	1
264	0.08	5	> 500	0.95	1	10
420	0.97	1	1	0.95	1	1
720	0.97	1	1	0.95	1	1
732	0.92	1	1	0.95	1	1

Table III— *Continued*

Obs	ρ^*	Critical L (t-statistic)	Critical L (R^2)	ρ^*	Critical L (t-statistic)	Critical L (R^2)
611	0.95	1	1	0.95	1	1
611	0.97	1	1	0.95	1	1
66	0.66	2	2	0.95	1	2
184	0.79	2	7	0.95	1	3
824	0.97	1	1	0.95	1	2
552	0.98	1	1	0.95	1	1

are examined, the more persistent ones are likely to be chosen, and the spurious regression problem is amplified. The slope coefficients are centered near zero, so the bias does not increase the average slopes of the selected regressors. Again, spurious regression works through the standard errors.

We can also say that spurious regression makes the data mining problem worse. For a given value of L, the critical t-ratios and R^2 values increase moving down the rows of Table III. For example, with $L = 250$ and true $R^2 = 0$, we can account for pure data mining with a critical t-ratio of 5.2. But when the true R-squared is 15 percent, the critical t-ratio rises to 6.3. The differences moving down the rows are even greater when $T = 824$, in Panel D. Thus, in the situations where the spurious regression bias is more severe, its impact on the data mining problem is amplified.

Finally, Panel E of Table III revisits the studies from the literature in view of spurious regression and data mining. We report critical values for L, the number of instruments mined, sufficient to render the regression t-ratios and R-squares insignificant at the five percent level. We use two assumptions about persistence in the true expected returns: (1) ρ^* is set equal to the sample values from the studies, as in Table I, or (2) $\rho^* = 95$ percent. With only one exception, the critical values of L are 10 or smaller. The exception is where the instrument is the lagged excess return on a two-month Treasury bill, following Campbell (1987). This is an interesting example because the instrument is not very autocorrelated, at 8 percent, and when we set $\rho^* = 8$ percent there is no spurious regression effect. The critical value of L exceeds 500. However, when we set $\rho^* = 95$ percent in this example, the critical value of L falls to 10, illustrating the strong interaction between the data mining and spurious regression effects.

IV. Conclusions

We study regression models where lagged variables predict stock returns, focusing on the issues of data mining and spurious regression. The spurious regression problem is related to the classic studies of Yule (1926) and Granger and Newbold (1974). Unlike the regressions in those papers, asset pricing regressions use rates of return, which are not highly persistent, as the dependent variables. However, asset returns are the expected returns plus unpredictable noise. If the *expected* returns are persistent, there is a risk of finding a spurious relation between the return and an independent, highly autocorrelated lagged variable.

When there is no persistence in the true expected return, the spurious regression phenomenon is not a concern. This is true even when the measured regressor is highly persistent. This implies that spurious regression is not a problem from the perspective of testing the null hypothesis that expected stock returns are unpredictable, even if a highly autocorrelated regressor is used. The evidence that expected stock returns vary over time is therefore not overturned by spurious regression bias.

Given persistent expected returns, we find that spurious regression can be a serious concern. The problem for stock returns gets worse as the autocorrelation in the expected return increases, and as the fraction of the stock return variance attributed to the conditional mean increases. Assuming that expected returns are as persistent as the median instrument in the samples of nine classic studies, we find that 7 of the 17 statistics that would be considered significant using traditional standards are no longer significant in view of the spurious regression bias. We therefore call into question the validity of specific instruments identified in the literature, such as the term spread, book-to-market ratio, and dividend yield.

Data mining, in the form of a search through the data for high-R^2 predictors, results in regressions whose apparent explanatory power occurs by chance. Consistent with Foster et al. (1997), if between 10 and 500 instruments are examined, depending on the study, all of the univariate regression results summarized in Table I become insignificant. In the presence of spurious regression, persistent variables are likely to be mined, and the two effects reinforce each other. As a result, the critical values needed for significant t-statistics and regression R-squares increase. If the expected return accounts for 10 percent of the stock return variance, mining among 5 to 10 instruments has as much impact as 50 to 100 instruments with no spurious regression. Assuming we sift through only 10 instruments, all of the regressions from the previous studies in Table I appear consistent with a spurious mining process.

Our results have distinct implications for tests of predictability and model selection. In tests of predictability, the researcher chooses a lagged instrument and regresses future returns on the instrument. The null hypothesis is that the slope coefficient is zero. Spurious regression presents no problem from this perspective, because under the null hypothesis, the expected return is not persistent. In model selection, the researcher chooses a lagged instrument to model time variation in expected returns, for purposes such as implementing or testing an asset pricing model. Here is where the spurious regression problem is the most pernicious.

The pattern of evidence for the instruments in the literature is similar to what is expected under a spurious mining process with an underlying persistent expected return. In this case we would expect instruments to arise, then fail to work out of sample. With fresh data, new instruments would arise, then fail. The dividend yield rose to prominence in the 1980s, but fails to work in post-1990 data (e.g., Goyal and Welch (2002) and Schwert (2002)). The book-to-market ratio seems to have weakened in recent data. With fresh data, new instruments appear to work (e.g., Lettau and Ludvigson (2001) and Lee, Myers, and Swaminathan, (1999)). There are two implications. First, we should be concerned that these

new instruments are likely to fail out of sample. Second, any stylized facts based on empirically motivated instruments and asset pricing tests based on such instruments should be viewed with skepticism.

Appendix: The Sample of 500 Instruments

All the data come from the web site Economagic.com: Economic Time Series Page, maintained by Ted Bos. The sample consists of all monthly series listed on the main homepage of the site, except under the headings of LIBOR, Australia, Bank of Japan, and Central Bank of Europe. From the Census Bureau, we exclude Building Permits by Region, State, and Metro Areas (more than 4,000 series). From the Bureau of Labor Statistics, we exclude all Noncivilian Labor Force data and State, City, and International Employment (more than 51,000 series). We use the Consumer Price Index (CPI) measures from the city average listings, but include no finer subcategories. The Producer Price Index (PPI) measures include the aggregates and the two–digit subcategories. From the Department of Energy, we exclude data in Section 10, the International Energy series.

We first randomly select (using a uniform distribution) 600 out of the 10,866 series that were left after the above exclusions. From these 600, we eliminated series that mixed quarterly and monthly data and extremely sparse series, and took the first 500 from what remained.

Because many of the data are reported in levels, we tested for unit roots using an augmented Dickey–Fuller test (with a zero order time polynomial). We could not reject the hypothesis of a unit root for 361 of the 500 series, and we replaced these series with their first differences.

We estimate a sample correlation matrix of the 500 instruments as follows. We take each pair of instruments and compute the sample correlation between the two series, using all of the periods in which our data for the two series overlap. For some pairs, there is no overlapping data. For these cases we substitute the average of all the sample correlations that we can compute.

REFERENCES

Bekaert, Geert, Robert J. Hodrick, and David Marshall, 1997, On biases in tests of the expectations hypothesis of the term structure, *Journal of Financial Economics* 44, 309–348.

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.

Boudoukh, Jacob, and Matthew Richardson, 1994, The statistics of long-horizon regressions, *Mathematical Finance* 4, 103–120.

Breen, William, Lawrence R. Glosten, and Ravi Jagannathan, 1989, Economic significance of predictable variations in stock index returns, *Journal of Finance* 44, 1177–1190.

Campbell, John Y., 1987, Stock returns and the term structure, *Journal of Financial Economics* 18, 373–400.

Campbell, John Y., and Robert Shiller, 1988, The dividend ratio and small sample bias, *Economics Letters* 29, 325–331.

Cochrane, John H., 1999, New facts in finance, *Economic Perspectives* 23, 36–58.

Conrad, Jennifer, and Gautam Kaul, 1988, Time variation in expected returns, *Journal of Business* 61, 409–425.

Cox, John C., Jonathan E. Ingersoll Jr., and Stephen A. Ross, 1985, A theory of the term structure of interest rates, *Econometrica* 53, 363–384.

Fama, Eugene F., 1970, Efficient capital markets: A review of theory and empirical work, *Journal of Finance* 25, 383–417.

Fama, Eugene F., 1990, Stock returns, expected returns, and real activity, *Journal of Finance* 45, 1089–1108.

Fama, Eugene F., and Kenneth R. French, 1988a, Dividend yields and expected stock returns, *Journal of Financial Economics* 22, 3–25.

Fama, Eugene F., and Kenneth R. French, 1988b, Permanent and temporary components of stock prices, *Journal of Political Economy* 96, 246–273.

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

Fama, Eugene F., and G. William Schwert, 1977, Asset returns and inflation, *Journal of Financial Economics* 5, 115–146.

Ferson, Wayne, and Campbell R. Harvey, 1991, Sources of predictability in portfolio returns, *Financial Analysts Journal* 3, 49–56.

Fleming, Jeff, Chris Kirby, and Barbara Ostdiek, 2001, The economic value of volatility timing, *Journal of Finance* 61, 329–352.

Foster, F. Douglas, Tom Smith, and Robert E. Whaley, 1997, Assessing goodness-of-fit of asset pricing models: The distribution of the maximal R-squared, *Journal of Finance* 52, 591–607.

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

Goyal, Amit, and Ivo Welch, 2003, Predicting the equity premium with dividend ratios, *Management Science* (forthcoming, May).

Granger, Clive W.J., and Paul Newbold, 1974, Spurious regressions in economics, *Journal of Econometrics* 4, 111–120.

Harvey, Campbell R., 1989, Time-varying conditional covariances in tests of asset pricing models, *Journal of Financial Economics* 24, 289–318.

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

Huberman, Gur, and Shmuel Kandel, 1990, Market efficiency and Value Line's record, *Journal of Business* 63, 187–216.

Kandel, Shmuel, and Robert F. Stambaugh, 1990, Expectations and volatility of consumption and asset returns, *Review of Financial Studies* 3, 207–232.

Kandel, Shmuel, and R.F. Stambaugh, 1996, On the predictability of stock returns: An asset-allocation perspective, *Journal of Finance* 51, 385–424.

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

Kendall, Maurice G., 1954, A note on the bias in the estimation of autocorrelation, *Biometrica* 41, 403–404.

Kim, Myung J., Charles R. Nelson, and Richard Startz, 1991, Mean reversion in stock prices? A reappraisal of the empirical evidence, *Review of Economic Studies* 58, 515–528.

Kothari, S.P., and Jay Shanken, 1997, Book-to-market time series analysis, *Journal of Financial Economics* 44, 169–203.

Lanne, Markku, 2002, Testing the predictability of stock returns, *Review of Economics and Statistics* 84, 407–415.

Lee, Charles, James Myers, and Bhaskaran Swaminathan, 1999, What is the intrinsic value of the Dow? *Journal of Finance*, 1693–1742.

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

Lo, Andrew W., and A.Craig MacKinlay, 1988, Stocks prices do not follow random walks, *Review of Financial Studies* 1, 41–66.

Lo, Andrew W., and A.Craig MacKinlay, 1990, Data snooping in tests of financial asset pricing models, *Review of Financial Studies* 3, 431–467.

Lucas, Robert E. Jr., 1978, Asset prices in an exchange economy, *Econometrica* 46, 1429–1445.

Marmol, Francesc, 1998, Spurious regression theory with nonstationary fractionally integrated processes, *Journal of Econometrics* 84, 233–250.

Merton, Robert C., 1973, An intertemporal capital asset pricing model, *Econometrica* 41, 867–887.

Nelson, Charles, and Myung J. Kim, 1993, Predictable stock returns: The role of small sample bias, *Journal of Finance* 48, 641–661.

Newey, Whitney K., and Kenneth D. West, 1987, A simple, positive definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica* 55, 703–708.

Pesaran, M.Hashem, and Allan Timmermann, 1995, Predictability of stock returns: Robustness and economic significance, *Journal of Finance* 50, 1201–1228.

Phillips, Peter C.B., 1986, Understanding spurious regressions in econometrics, *Journal of Econometrics* 33, 311–340.

Phillips, Peter C.B., 1998, New tools for understanding spurious regressions, *Econometrica* 66, 1299–1326.

Pontiff, Jeffrey, and Lawrence Schall, 1998, Book-to-market as a predictor of market returns, *Journal of Financial Economics* 49, 141–60.

Schwert, G. William, 2002, Anomalies and market efficiency, in George M. Constantinides, Milton Harris, René M. Stulz, eds.: *Handbook of the Economics of Finance*, North Holland: Amsterdam.

Simin, Timothy, 2002, The (poor) predictive performance of asset pricing models, Working paper, Pennsylvania State University.

Stambaugh, Robert S., 1999, Predictive regressions, *Journal of Financial Economics* 54, 315–421.

Valkanov, Rossen, 2003, Long-horizon regressions: Theoretical results and applications, *Journal of Financial Economics* 68, 201–232.

Yule, George U., 1926, Why do we sometimes get nonsense correlations between time series? A study in sampling and the nature of time series, *Journal of the Royal Statistical Society* 89, 1–64.