

# TESTING THE PREDICTABILITY OF STOCK RETURNS

Markku Lanne\*

*Abstract*—Previous literature indicates that stock returns are predictable by several strongly autocorrelated forecasting variables, especially at longer horizons. It is suggested that this finding is spurious and follows from a neglected near unit root problem. Instead of the commonly used  $t$ -test, we propose a test that can be considered as a general test of whether the return can be predicted by any highly persistent variable. Using this test, no predictability is found for U.S. stock return data from the period 1928–1996. Simulation experiments show that the standard  $t$ -test clearly overrejects whereas our proposed test controls size much better.

## I. Introduction

SEVERAL empirical studies have documented the predictability of stock returns by various forecasting variables. Among the most successful forecasting variables are financial variables (such as the dividend-price ratio and term structure variables) and macroeconomic fundamentals (such as output and inflation). Specifically, the recent literature has found predictability at long return horizons. (See the exhaustive survey by Kaul (1994).) Typically, empirical work in this field has relied on regression models in which the  $k$ -period stock return is regressed on one or more of these variables, and significant  $t$ - or  $F$ -tests are interpreted in favor of the predictability of stock returns. The purpose of this paper is to address the econometric problems that are involved in testing for predictability in the simple regression model framework.

Originally, tests of the predictability of asset returns were motivated by market efficiency; that is, it was claimed that predictability would be inconsistent with the efficient markets paradigm. Nowadays, the view is that the predictability of time-varying expected returns can exist in efficient markets, and the interesting question has rather become how much predictability is compatible with efficiency under different asset-pricing models. (See, for example, Kirby (1998).) Still, the question of whether long-horizon returns are more predictable than short-horizon returns remains unresolved. At practical level, the issue is important because long-horizon predictability could be exploited in investment strategies based on dynamic asset allocation models.

The econometric problems of the regression models used to examine return predictability have been recognized already early in the literature. There are two main difficulties. First, all the typical forecasting variables are highly persistent, having both a large autoregressive root in their univariate representation and not really exogenous but lagged endogenous (predetermined) variables. These two facts to-

gether result in high simultaneity in the system consisting of a univariate model for the forecasting variable and the regression model. It was shown already by Mankiw and Shapiro (1986), and Stambaugh (1986) that in this case the slope coefficient can be substantially biased. Furthermore, the recent results of Elliott and Stock (1994) indicate that the standard normal distribution may not be a good approximation to the true null distribution of the  $t$ -statistic on the slope coefficient even asymptotically, and specifically, employing critical values from that distribution can lead to sizeable overrejection. The second problem follows from the use of overlapping data in long-horizon regressions. This causes the error term to be serially correlated, and the estimators of the autocorrelation consistent standard errors for the coefficients are known to be poorly behaved in finite samples (Richardson and Stock, 1989). In fact, the standard errors tend to be underestimated, which enforces the overrejection problem. (See, for example, Nelson and Kim (1993) and the references therein.) Given these two problems, it is plausible that predictability is often spuriously found when standard methods are employed. Recently, the overrejection problem in long-horizon returns was demonstrated by the Monte Carlo simulation experiments of Kirby (1997), who predominantly attributed the result to the second problem, though.

A few recent attempts to resolve the econometric problems in testing for long-horizon predictability are based on simulation methods (bootstrap and its modifications). Hodrick's (1992) simulations revealed overrejection in a model in which the dividend yield is a unit root process. He still concluded that dividend yields have predictive power for stock returns in the U.S. market. Nelson and Kim (1993) found some predictability of postwar U.S. stock returns using randomized (bootstrapped without replacement) null distributions for the test statistics. They also found clear differences between the simulated and asymptotic null distributions and, hence, strongly recommended simulation methods. Employing bootstrap methods, Goetzmann and Jorion (1993), on the other hand, failed to find any strong evidence for predictability of the U.S. stock returns by dividend yields in the period 1927–1990. Recent results in the econometric literature, however, indicate that the bootstrap fails to provide an asymptotically valid method of inference in regression models in which the regressor has a near unit root. Thus, all the procedures based on simulation are likely to be unreliable here. This is also true of the likewise computer-intensive subsampling method that was used by Wolf (2000) to attack the problem. This method is, however, expected to yield improved small-sample inference over the standard  $t$ -test when the regressor has a fixed root close to but strictly less than unity. If the exact unit root case could be excluded, this method should be far more

Received for publication October 22, 1999. Revision accepted for publication April 2, 2001.

\* University of Helsinki.

I would like to thank Pentti Saikkonen, John Y. Campbell (co-editor), and two anonymous referees for useful comments. The usual disclaimer applies. This paper is a part of the research program of the Research Unit on Economic Structures and Growth (RUESG) in the Department of Economics of the University of Helsinki. Financial support from the Yrjö Jahansson Foundation is gratefully acknowledged.

reliable than the methods previously mentioned, and Wolf did not find any predictability of the U.S. stock return (1947–1995) by dividend yields.

Under the alternative hypothesis that the return is predictable by some highly persistent forecasting variable, the return can be characterized as a near unit root plus noise process, whereas under the null hypothesis it is a strictly stationary process. Thus, testing the predictability of the returns by such a forecasting variable is equivalent to testing the stationarity of the returns. If the returns are stationary, they cannot be predicted by forecasting variables with a large, possibly unit autoregressive root. So, instead of actually running the regressions of the returns at different horizons on several forecasting variables, we suggest a stationarity test (moving-average unit root test) of the return series. Contrary to a conventional  $t$ -test, this test has correct size asymptotically, and it is shown in the appendix that it also has good power against local alternatives. Furthermore, it has the additional advantage of being a general test of predictability because it does not use information on any actual forecasting variable. Using this test, we do not find any predictability of the U.S. stock returns in the 1928–1996 period. Moreover, small-sample simulation experiments indicate that the test controls size well.

The plan of the paper is as follows. Section II discusses the problems of standard inference in long-horizon stock return regressions and presents the MA unit root test as a potential solution. Section III presents the empirical results, and section IV concludes.

## II. Econometric Methodology

In this section, we examine the potential problems with the econometric methodology commonly used in testing for the predictive ability of different forecasting variables for stock returns and suggest an alternative testing procedure. Typically, returns, measured over various horizons, are regressed on a variable such as the dividend-price ratio or some interest rate variable, and the significance of the slope coefficient is tested with a  $t$ -test. Often, the candidate forecasting variables are strongly autocorrelated and their stationarity cannot be assumed. Furthermore, they are usually only predetermined but not exogenous. These facts induce bias in the slope coefficient of the forecasting regression equation, and this bias has been seen as a major econometric problem in testing for predictability (Nelson and Kim, 1993). The recent results of Elliott and Stock (1994), however, suggest that standard inference is not, in general, even asymptotically valid in a case like this when the order of integration of the regressor is unknown. Another finite sample problem discussed in the previous literature is caused by using overlapping data, which makes the error term of the regression model autocorrelated and thus complicates the estimation of standard errors (Richardson and Stock, 1989).

To see the near unit root problem, consider the following simple forecasting regression typically used in empirical studies,

$$r_{t+1} + \dots + r_{t+k} = \mu_r + \beta x_t + u_{2t+k},$$

where  $r_t + \dots + r_{t+k-1}$  is the real continuously compounded  $k$ -period stock return and  $x_t$  is some forecasting variable. Denoting the  $k$ -period return on the left-hand side by  $r_{t+k}^k$  the model can equivalently be written as

$$r_t^k = \mu_r + \beta x_{t-k} + u_{2t}. \quad (1)$$

Assume that  $x_t$  has an autoregressive representation,

$$x_t = \mu_x + \rho x_{t-1} + u_{1t}, \quad a(L)u_{1t} = \epsilon_{1t}, \quad (2)$$

where  $(1 - \rho L)a(L)$  is a  $p$ th-order lag polynomial with the largest characteristic root equal to  $\rho$  and all roots of  $a(L)$  outside the unit circle. To parametrize the high persistence of  $x_t$ , we consider the sequence  $\rho = 1 + c/T$ , where  $c$  is a constant and  $T$  is the sample size. Thus, if  $c = 0$ , the process has a unit root, while if  $c < 0$  the process is stationary. When overlapping data is used to estimate the parameters of model (1), the error term  $u_{2t}$  is serially correlated with nonzero autocorrelation coefficients up to lag  $k - 1$  even if the specification is correct. Thus, it can be assumed that  $u_{2t}$  follows an MA( $k - 1$ ) process; that is,  $u_{2t} = b(L)\epsilon_{2t}$ , where  $b(L)$  is a  $(k - 1)$ th-order lag polynomial with all the characteristic roots outside the unit circle. Let us further assume that the error term  $\epsilon_t = (\epsilon_{1t}, \epsilon_{2t})'$  is a martingale difference sequence with constant conditional covariance matrix  $E(\epsilon_t | \epsilon_{t-1}, \epsilon_{t-2}, \dots) = \Sigma$ ,  $\sup_t E\epsilon_{it}^4 < \infty$ ,  $i, j = 1, 2$ , and  $E\epsilon_{10}^4 < \infty$ . Let  $\Omega = [\omega_{ij}]$ ,  $i, j = 1, 2$  denote the long-run covariance matrix of the error terms  $u_t = (u_{1t}, u_{2t})'$  ( $2\pi$  times their spectral density matrix at frequency zero). Finally, assume that

$$\begin{aligned} T^{-1/2} \sum_{t=1}^{[Ts]} (u_{1t}/\omega_{11}, u_{2t}/\omega_{22}) &\Rightarrow (W_1(s), W_2(s))' \\ &= W(s), \end{aligned}$$

where  $\Rightarrow$  denotes weak convergence on  $D[0, 1]$ ,  $W(s)$  is a two-dimensional Brownian motion with covariance matrix  $\bar{\Omega} = [\bar{\omega}_{ij}]$  with  $\bar{\omega}_{11} = \bar{\omega}_{22} = 1$ , and  $\bar{\omega}_{12} = \bar{\omega}_{21} = \delta = \omega_{12}/(\omega_{11}\omega_{22})^{1/2}$ .

Under these assumptions, using an autocorrelation and heteroskedasticity consistent standard error, the  $t$ -statistic,  $t_\beta$ , testing the null hypothesis  $\beta = 0$  in equation (1) has the following asymptotic null distribution (Elliott and Stock, 1994, result (5))<sup>1</sup>

<sup>1</sup> Elliott and Stock (1994) considered a system in which the regressor was  $x_{t-1}$  instead of  $x_{t-k}$ . Asymptotically, this makes no difference because  $x_{t-k}$  can always be written as  $x_{t-1} - \Delta x_{t-1} - \dots - \Delta x_{t-k+1}$  and it can

$$t_\beta \Rightarrow \delta \left( \int_0^1 J_c^\mu(s) ds \right)^{-1/2} \int J_c^\mu dW_1(s) + (1 - \delta^2)z, \tag{3}$$

where  $J_c^\mu(s) = J_c(s) - \int_0^1 J_c(r)dr$  is a demeaned Ornstein-Uhlenbeck process with  $J_c(s)$  generated by the stochastic differential equation  $dJ_c(s) = cJ_c(s)ds + dW_1(s)$  with  $J_c(0) = 0$ , and  $z$  is a standard normal random variable distributed independently of  $W(s)$  and  $J_c(s)$ .

This result shows that inference about  $\beta$  relying on the standard normal distribution is not asymptotically justified if  $\rho$  is large and there is simultaneity in the system; that is,  $\delta$  is not zero. In this case, the asymptotic null distribution is nonstandard, and the size distortions from using the normal approximation are the larger the bigger  $\delta$  is in absolute value. As  $\rho$  becomes smaller (or, equivalently,  $-c$  becomes larger), the limiting distribution gets closer to the standard normal distribution; in the limit  $c \rightarrow -\infty$ ,  $t_\beta$  is distributed as a standard normal variable (Nabeya and Sørensen, 1994). For fixed  $c$ , when  $\delta < 0$ , the limiting distribution in equation (3) lies to the right, and, when  $\delta > 0$ , to the left of the standard normal distribution. Hence, the standard two-sided  $t$ -test tends to overreject when critical values from the standard normal distribution are used.<sup>2</sup> This result offers one potential explanation for the finding that stock returns seem to be predictable by various persistent variables.

It is obvious from equation (3) that the asymptotic distribution of  $t_\beta$  depends on the values of the two nuisance parameters,  $\delta$  and  $c$ . If these could be consistently estimated, then the appropriate critical values could in each application be computed. Although  $\delta$  indeed is consistently estimated by any consistent estimator of the long-run correlation between  $u_{1t}$  and  $u_{2t}$ ,  $c$  is not consistently estimable, and so this approach is not applicable. One might conjecture that a consistent estimator of  $c$  is given by  $T(\hat{\rho} - 1)$ , where  $\hat{\rho}$  is the OLS estimator of  $\rho$ . This, however, is not true because (Phillips, 1987, Theorem 1)

$$T(\hat{\rho} - 1) = T(\hat{\rho} - \rho) + T(\rho - 1) \Rightarrow \int_0^1 J_c^\mu(s) dW(s) \times \left( \int_0^1 J_c^\mu(s)^2 ds \right)^{-1} + c. \tag{4}$$

Thus,  $T(\hat{\rho} - 1)$  converges to a distribution dependent on the true value of the local-to-unity parameter,  $c$ , which is

be shown that the terms involving the differences in the  $t$ -statistic are  $o_p(1)$ .

<sup>2</sup>Pretesting for a unit root in the regressor does not solve the problem either, see Elliott and Stock (1994).

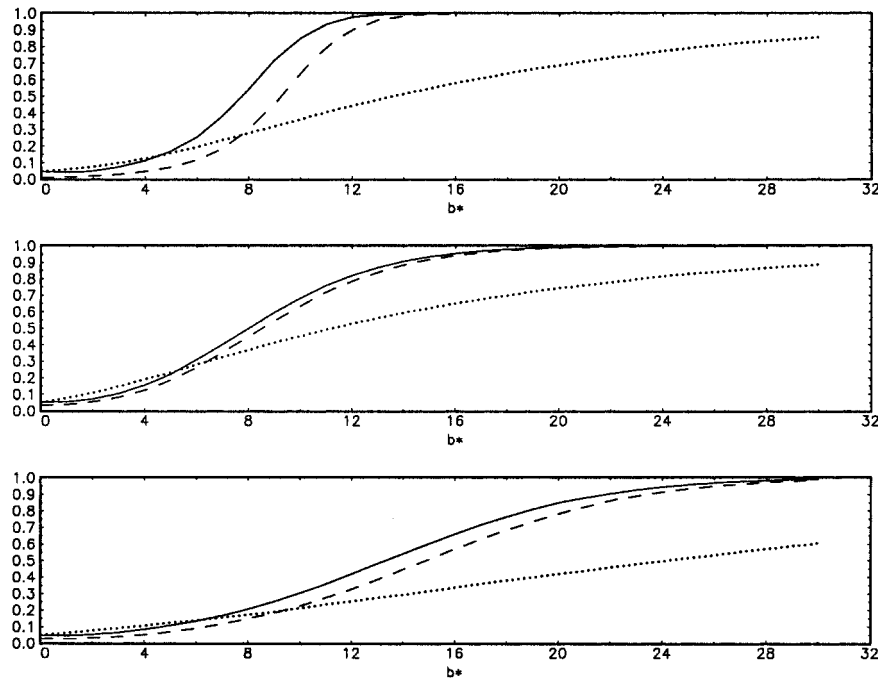
unknown. Therefore, the value that this estimator gives is nothing but a random draw from the preceding distribution. This also invalidates the use of bootstrap methods in this kind of models, as was recently pointed out by Stock (1997). Because these methods depend on the unavailable consistent estimate of  $c$ , a model estimated from data and used for generating bootstrap samples fails to deliver the true null distribution of, say,  $t_\beta$ .<sup>3</sup> Therefore, the empirical results on the predictability of stock returns due to Hodrick (1992), Goetzmann and Jorion (1993), and Nelson and Kim (1993) obtained using bootstrap and closely related approaches are not theoretically justified.

There are by now several asymptotically valid procedures of inference in simple regression models when the order of integration of the regressor is unknown; these are surveyed by Stock (1997). Because of the error autocorrelation caused by overlap in equation (1), the approach based on stationarity (moving-average unit root) tests seems to be the most practical here. The idea of this test is as follows. Under the null hypothesis that  $\beta = 0$ ,  $r_t^k$  is a stationary MA( $k - 1$ ) process, whereas under the alternative it is the sum of that process and the process  $x_t$  with a root local to unity. Hence, the test of the null hypothesis can be based on testing whether  $r_t^k$  is stationary. It is worth noting that this procedure does not explicitly use information on the regressor, so that it can be interpreted as testing whether the  $k$ -period stock returns can be predicted by any near unit root process,  $x_t$ . Of course, the power of the test depends on the long-run correlation between the error terms,  $u_{1t}$  and  $u_{2t}$ . This approach was originally suggested by Wright (2000) for the closely related problem of forming confidence intervals for the coefficients of a cointegration vector. In that context, the idea is to invert the stationarity test to find the confidence set of the cointegrating coefficients such that the cointegration vector is stationary. The approach has the advantage of being valid even if the variables do not have exact unit roots.

Several alternative stationarity tests (that is, tests for an MA unit root) are available. In model (1), the tests of Nyblom and Mäkeläinen (1983), Kwiatkowski et al. (1992), and Saikkonen and Luukkonen (1993) are equivalent when the error term is not autocorrelated. Complications do, however, arise in cases in which the order of autocorrelation of the error term is high. The test suggested by Saikkonen and Luukkonen involves estimating the parameters of this MA( $k - 1$ ) model for the error term. Because  $k$  can be very large in long-horizon regressions, this approach is computationally burdensome and probably infeasible here in

<sup>3</sup>The asymptotic invalidity of the bootstrap has been proved in at least two closely related cases. Stock and Watson (1996) showed that the bootstrap does not control size in inference on the parameters of the cointegrating vector when the regressors do not necessarily have a unit root. Basawa et al. (1991), on the other hand, demonstrated that, in the special case of the AR(1) model with a unit root, the bootstrap distribution of the slope coefficient converges to a random distribution, even if the error distribution is known to be normal.

FIGURE 1.—LOCAL ASYMPTOTIC POWER OF THE INFEASIBLE  $t$  (SOLID LINE), SUP- $c$  (DASHES), AND  $L^\mu$  (DOTS) TESTS OF THE NULL HYPOTHESIS  $\beta = 0$  AGAINST  $\beta = b/T$  AT THE 5% LEVEL FOR SOME VALUES OF  $c$  AND  $\delta$



(Top:  $c = -5$ ,  $\delta = -0.9$ ; middle:  $c = -5$ ,  $\delta = -0.3$ ; bottom:  $c = -20$ ,  $\delta = -0.3$ );  $b^* = (\omega_{11}/\omega_{22})^{1/2}b$ .

practice. Therefore, the test due to Nyblom and Mäkeläinen, and Kwiatkowski et al. (hereafter, KPSS test) will be employed in the empirical part of the paper. The test statistic is

$$L^\mu = \hat{\sigma}_{u_2}^{-2} T^{-2} \sum_{t=1}^T \left( \sum_{s=1}^t (r_s^k - \bar{r}^k) \right)^2, \quad (5)$$

where  $\bar{r}^k$  is the sample mean of  $r^k$  and  $\hat{\sigma}_{u_2}^2$  is a consistent estimator of the long-run variance of the error term,  $u_{2t}$ . Kwiatkowski et al. show that, under the null of stationarity ( $\beta = 0$  here), equation (5) converges to a functional of the standard Brownian bridge  $V(s)$ ,

$$L^\mu \Rightarrow \int_0^1 V(s)^2 ds.$$

Although this test was developed for testing for stationarity against an exact unit root, it also has power against near unit root alternatives. It is shown in the appendix that the test is consistent (against fixed alternatives  $\beta \neq 0$ ) for the null hypothesis  $\beta = 0$  in equation (1) if  $\sigma_{u_2}^2$  is estimated using the Bartlett kernel to be employed in the empirical part of the paper, provided the lag truncation parameter is  $o(T)$ . The local asymptotic power calculations presented in the appendix indicate that the power rapidly declines as the forecasting variable  $x_t$  becomes “more stationary” so that the test is indeed useful only if  $x_t$  can really be expected to be strongly autocorrelated.

Although the  $L^\mu$  test can be shown to be consistent using the local-to-unity asymptotics, it is possible to construct examples in which, for a fixed level of persistence of the expected stock return, the realized return is white noise, and hence the test has zero power. Such a case would arise, for instance, if the expected stock return followed a persistent AR(1) process, but the realized return followed an ARMA(1,1) process having cancelling roots, with the MA part stemming from news about future dividends (cf. Campbell, Lo, and MacKinlay (1997), section 7.1). However, as the persistence of the expected stock return increases, the variance of its innovation must decrease in relation to the variance of the innovation to future dividends in order to keep the realized return process uncorrelated. This, in turn, implies that the standard  $t$ -test can also have very low power in this case.

Instead of the approach based on stationarity tests, one of the asymptotically conservative tests due to Cavanagh, Elliott, and Stock (1995) could be employed. Although the sup- $c$  and Bonferroni tests are applicable, the test based on Scheffe-type confidence intervals would become very complicated in practice because of the overlap in the forecasting regression at longer horizons. Both the sup- $c$  and Bonferroni tests are based on the usual  $t$ -test statistic with critical values computed from equation (3) corresponding to an estimate of  $\delta$  and a range of values of  $c$ . The idea is to find the minimum and maximum of the critical values in this range, and the null hypothesis is rejected if the value of the test statistic lies below the minimum or above the maximum

TABLE 1.—90%-CONFIDENCE INTERVALS FOR THE LARGEST AUTOREGRESSIVE ROOT OF THE LOG DIVIDEND-PRICE RATIO

Period	90%-Confidence Interval
1928–1996	0.969, 1.002
1928–1951	0.912, 1.003
1952–1996	0.966, 1.006

The confidence intervals are computed by inverting augmented Dickey-Fuller statistics (Stock, 1991).

at a given significance level. In the Bonferroni test, the range of  $c$  to consider is first estimated by inverting a unit root test of the regressor, whereas in the sup- $c$  test the range covers all the possible values of the local-to-unity parameter,  $c$ . In practice, we may want to exclude explosive roots, that is, consider only cases in which  $c \leq 0$ . The calculations of Cavanagh et al. (1995) show that, even with size adjustments, these tests, especially the sup- $c$  test, are rather conservative.

Information on the loss in local asymptotic power incurred by using the  $L^\mu$  or sup- $c$  test because  $c$  is unknown is provided by the local asymptotic power curves of these tests and the infeasible  $t$ -test with known  $c$  in figure 1.<sup>4</sup> The asymptotic power loss of the  $L^\mu$  test can be very large, especially if  $\delta$  is large in absolute value. The local asymptotic power of the  $L^\mu$  test also decreases with the value of  $c$ , as expected. Hence, although the test is consistent for any value of  $c$ , it is advisable to use it only when the forecasting variable is expected to be strongly autocorrelated. The power loss of the sup- $c$  test is much smaller, but, as the finite-sample simulation experiments in section III indicate, it also tends to overreject when the error term of the forecasting regression is autocorrelated whereas such over-rejection is not a serious problem with the  $L^\mu$  test.

### III. Empirical Results

In this section, the predictability of the U.S. stock returns is examined. The prices and dividends are computed from the monthly value-weighted CRSP index of stocks traded on the NYSE, AMEX, and NASDAQ, along the lines of Hodrick (1992). The dividend-price ratio is measured as the sum of the dividends paid on the index over the previous twelve months, divided by the current level of the index. The real returns are computed using the CPI (for all urban consumers) data from the Bureau of Labor Statistics. In the following empirical analysis, the logarithm of the dividend-price ratio is used instead of the level. Following previous literature, the results are presented, in addition to the entire sample 1928–1996 (828 observations), also separately for the 1928–1951 (288 observations) and 1952–1996 (540 observations) subsample periods.

The 90%-confidence intervals for the largest autoregressive roots  $\rho$  of the log dividend-price ratio for the different periods are presented in table 1. Unity is included in all the

<sup>4</sup> The calculations of Cavanagh et al. (1995) suggest that the local asymptotic power properties of the Bonferroni test are not, in general, that much different from those of the sup- $c$  test.

TABLE 2.—ESTIMATES OF THE LONG-RUN CORRELATION BETWEEN THE ERROR TERMS OF THE REGRESSION OF THE  $k$ -PERIOD STOCK RETURN ON THE LOG DIVIDEND-PRICE RATIO, AND THE AUTOREGRESSION FOR THE LOG DIVIDEND-PRICE RATIO

$k$	1928–1996	1928–1951	1952–1996
1	−0.843	−0.888	−0.946
3	−0.863	−0.858	−0.902
6	−0.828	−0.801	−0.858
12	−0.797	−0.744	−0.751
24	−0.666	−0.525	−0.377
36	−0.566	−0.460	−0.287
48	−0.482	−0.420	−0.303

The estimates are computed using the Bartlett kernel estimator with automatic bandwidth selection (Newey and West, 1994).

intervals. Although the local-to-unity parameter,  $c$ , is not consistently estimable, we can get some idea of the relevance of the near unit root problem by computing the implied values corresponding to the lower limits of the confidence intervals. In the entire sample, this value is approximately  $-25$ , and, in the subsamples, it is  $-27$  and  $-17$ , respectively. These values indicate that there may be a near unit root problem in regression model (1). It also seems that the returns are more persistent in the latter subsample period, suggesting potentially even bigger problems for standard inference in that period. The higher persistence in the postwar period was also documented by Kim, Nelson, and Startz (1991). Whether the near unit root problem really exists depends on the simultaneity in the system consisting of the regression model and the univariate model for the log dividend-price ratio. This is examined in table 2, which presents the estimates of the long-run correlation coefficients of the error terms of equations (1) and (2) for the different sample periods and horizons. Simultaneity is high in every case, confirming the need for robust inference.<sup>5</sup>

<sup>5</sup> If the near unit root problem were the only reason for the potential near unit root problem, one would (at least asymptotically) expect the simultaneity to be higher in the regressions with longer forecast horizons. The

TABLE 3.—ESTIMATES OF THE SLOPE COEFFICIENTS AND THEIR  $t$ -STATISTICS IN THE REGRESSION OF THE  $k$ -PERIOD STOCK RETURN ON THE LOG DIVIDEND-PRICE RATIO

$k$	1928–1996	1928–1951	1952–1996
1	0.009 (0.679)	0.015 (0.481)	0.016 (1.371)
3	0.034 (1.352)	0.061 (1.029)	0.057 (2.246)
6	0.066 (1.426)	0.097 (1.046)	0.135 (2.645)
12	0.163 (2.193)	0.284 (1.991)	0.296 (3.019)*
24	0.341 (3.550)**	0.643 (4.033)**	0.530 (2.736)*
36	0.465 (3.826)**	0.883 (3.233)**	0.613 (2.628)*
48	0.595 (3.964)**	1.040 (4.539)**	0.713 (3.093)**

The  $t$ -statistics in the parentheses are computed using the Bartlett kernel estimator with automatic bandwidth selection (Newey and West, 1994).

\* and \*\* denote rejections in the sup- $c$  test at the 5% and 1% levels, respectively.

TABLE 4.—THE VALUES OF THE KPSS STATISTIC TESTING THE NULL HYPOTHESIS THAT THE SLOPE COEFFICIENT EQUALS ZERO IN THE REGRESSION OF THE  $k$ -PERIOD STOCK RETURN ON THE LOG DIVIDEND-PRICE RATIO

$k$	1928–1996	1928–1951	1952–1996
1	0.056	0.085	0.162
3	0.057	0.098	0.150
6	0.053	0.104	0.156
12	0.066	0.115	0.199
24	0.093	0.167	0.246
36	0.128	0.177	0.209
48	0.139	0.265	0.172

The critical values are 0.347, 0.463, and 0.739 at the 10%, 5%, and 1% levels, respectively. The long-run variance of the error term is estimated using the Bartlett kernel estimator with automatic bandwidth selection (Newey and West, 1994).

The slope coefficients and their  $t$ -statistics for the forecasting equation (1) are presented in table 3. The results resemble those in the previous literature: the estimates are positive, and, employing critical values from the standard normal distribution, with the exception of the shortest horizons we reject the null that the log dividend-price ratio does not predict future stock returns at the 5% level. The  $t$ -statistics use standard errors computed by the Bartlett kernel estimator with the bandwidth selected by the automatic procedure of Newey and West (1994).

For comparison, the sup- $c$  test was also considered. In the 5%-level test, the critical values were the 2.5% quantile from the standard normal distribution ( $-1.96$ ) and the 97.5% quantile from asymptotic null distribution (3) with  $\delta$  replaced by its estimated value and  $c = 0$ . The Bonferroni test would use tighter acceptance regions, and hence reject at least in all the cases for which the sup- $c$  test rejects. The test rejected at the 5% level for forecast horizons 24, 36, and 48 in all the sample periods. Although this test is asymptotically conservative, the finite-sample simulation experiments (following) show that it can severely overreject.

The results of the KPSS test are presented in table 4. The long-run variance of the error term is estimated using the Bartlett kernel estimator with the automatic bandwidth selection procedure of Newey and West (1994). As was mentioned, the consistency of the test requires the bandwidth to be  $o(T)$  both under the null and alternative hypotheses, which is guaranteed by this procedure. (See Hobijn, Franses, and Ooms (1998, Lemma 2).)<sup>6</sup> The test seems to be rather robust with regards to the lag length, though. In no case is the null hypothesis rejected at the 5% level. Hence, these results suggest that the log dividend-price ratio does not predict future stock returns even for longer horizons. Recalling that the test does not use information on the

opposite conclusion from table 2 suggests that inference in finite samples is presumably also affected by the strong autocorrelation in the error term when the forecast horizon is long.

<sup>6</sup> In fact, the bandwidth parameter is  $o(T^{13/27})$  in this case. Hobijn et al. (1998) also showed that the use of this procedure improves the small-sample properties of the KPSS test. Note that, for instance, in the lag selection scheme due to Andrews (1991), the bandwidth parameter is  $O(T)$  under the stochastic trend alternative. (See Choi (1994).) Of course, consistency could in this case be achieved by imposing a deterministic upper bound for the bandwidth parameter.

TABLE 5.—REJECTION RATES AT THE 5% NOMINAL LEVEL FOR THE  $t$ -, sup- $c$ , AND KPSS TESTS FOR THE SIGNIFICANCE OF THE LOG DIVIDEND-PRICE RATIO IN THE SIMPLE LINEAR REGRESSION MODEL FOR THE  $k$ -PERIOD STOCK RETURN WITH FIXED SAMPLE SIZE

$k$	$t$ -test	sup- $c$ test	KPSS test
1	0.197	0.038	0.077
3	0.209	0.043	0.073
6	0.227	0.051	0.072
12	0.262	0.067	0.075
24	0.314	0.114	0.081
36	0.357	0.159	0.084
48	0.401	0.199	0.097

The figures are based on 10,000 replications of a restricted VAR for the stock returns and log dividend-price ratios (with the coefficients of the lagged log dividend-price ratio restricted to zero in the equation for the stock return). The VAR model was estimated using data from the latter subsample period (1952–1996, 540 observations). The sample size is 540. The standard errors and long-run variances of the error term are estimated using the Bartlett kernel estimator with automatic bandwidth selection (Newey and West, 1994).

log dividend-price ratio itself, but can be interpreted as a test of whether stock returns are predictable by any near- $I(1)$  variable, it can be concluded that the stock returns cannot be forecast by any such variables.

To assess the finite-sample performance of the  $t$ -sup- $c'$  and KPSS tests, a small Monte Carlo simulation experiment was conducted. First, a restricted vector autoregression for the stock returns and log dividend-price ratios was estimated with the coefficient of the lag of the log dividend-price ratio restricted to zero in the equation for the stock return. Then, 10,000 samples of artificial data consisting of 540 observations were generated from this model, and the test statistics were computed. Following Hodrick (1992), the model was estimated with data from the latter subsample period, which seems to be the longest homogenous sample period available in view of the previous results concerning persistence. The Akaike information criterion selected a VAR(1) model, and diagnostic tests did not detect any ARCH effects, so homoskedasticity was assumed in the experiments and the error terms were generated from the normal distribution with mean zero and covariance matrix estimated from the residuals of the restricted VAR.

The empirical sizes of the  $t$ -, sup- $c$ , and KPSS tests are reported in table 5. The  $t$ -test clearly overrejects for all return horizons, and with longer horizons the problem gets worse. Somewhat surprisingly, the sup- $c$  test also seems to overreject at the 24-, 36-, and 48-month horizons. Because the sup- $c$  test is asymptotically valid (whereas the  $t$ -test is not), the differences in the rejection rates in the presence of high-order autocorrelation in the error term may be interpreted as measuring the extent of the problems of estimating the autocorrelation consistent standard errors. When the overlap is modest, the sup- $c$  test has approximately the correct size, but overrejections prevail at longer horizons with more-severe overlap. Although the KPSS test, in general, controls size relatively well, there is slight overrejection at longer forecast horizons, which can probably be ascribed to the well-known problems of estimating the long-run variance in finite samples. These results thus indicate that the KPSS test is the only reliable test of the three.

TABLE 6.—REJECTION RATES AT THE 5% NOMINAL LEVEL FOR THE  $t$ -, sup- $c$ , AND KPSS TESTS FOR THE SIGNIFICANCE OF THE LOG DIVIDEND-PRICE RATIO IN THE SIMPLE LINEAR REGRESSION MODEL FOR THE  $k$ -PERIOD STOCK RETURN WITH SAMPLE SIZE DEPENDENT ON  $k$

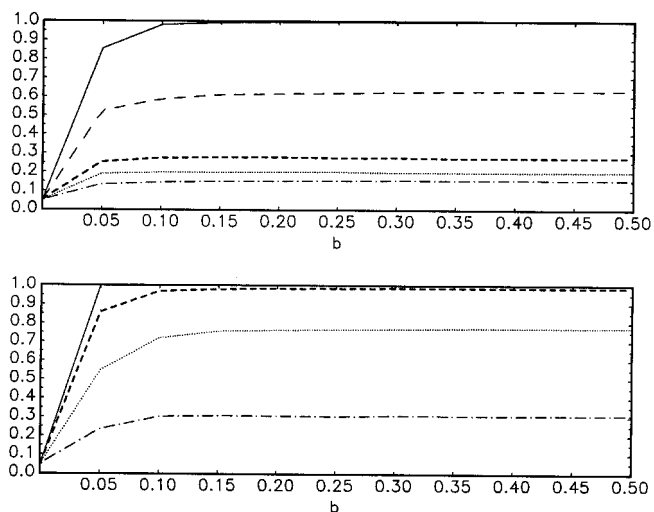
$k$	$t$ -test	sup- $c$ test	KPSS test
1	0.200	0.041	0.070
3	0.117	0.022	0.055
6	0.092	0.033	0.055
12	0.082	0.044	0.060
24	0.116	0.060	0.056
36	0.109	0.057	0.055
48	0.182	0.082	0.064

See the notes to table 5. The sample size is 540k, and the number of replications is 1,000.

The observation that the overrejection problem in the  $t$ -test gets more severe with longer forecast horizons agrees with the empirical results presented here and in the previous literature. This finding is not, however, completely in accordance with the figures in table 2, which suggest that the overrejection should be worse at short horizons. One plausible explanation is that the error autocorrelation increasing with the forecast horizon contributes to the overrejection along with the near unit root problem. To consider the relative contributions of the near unit root problem and observation overlap, a further simulation experiment was conducted. Instead of keeping the sample size fixed, we took the number of nonoverlapping observations as the relevant sample size in the spirit of Richardson and Stock (1989) and simulated samples with size 540 times the forecast horizon. The results are presented in table 6. Because of the computational burden, the figures are based on only 1,000 replications. Compared to the rejection rates in table 5, there is much less variation across the different horizons, especially as far as the  $t$ - and sup- $c$  tests are concerned. For the  $t$ -test, the rejection rate is highest for  $k = 1$ , and then it first declines with growing  $k$ , and for  $k \geq 24$  starts increasing. Likewise, the sup- $c$  test only overrejects for  $k \geq 24$ . These patterns are consistent with the explanation that, with growing forecast horizon, the near unit root problem is relieved but simultaneously the overlap gets worse, to the extent that at the longest horizons the extra observations are not sufficient to compensate for error autocorrelation.

The local asymptotic power calculations presented in figure 1 suggest that the KPSS test may lack power. To examine the finite-sample power properties, a further simulation experiment was conducted. The data-generating processes (DGP) were again estimated using the 540 last observations of the data set. In this experiment, the DGP of the regressor should remain the same irrespective of which alternative DGP for the stock return is considered, and for the log dividend-price ratio an AR(2) model was deemed sufficient by diagnostic tests. The stock returns were generated from an estimated AR(1) model with the lagged log dividend-price ratio multiplied by  $b$  ( $b = 0.05, 0.10, \dots, 0.50$ ) added. The power of the tests is expected to increase with the value of  $b$ . The error terms were generated from a

FIGURE 2.—EMPIRICAL REJECTION FREQUENCIES OF THE “SIZE-ADJUSTED” 5% LEVEL KPSS (TOP) AND INFEASIBLE  $t$ -TESTS (BOTTOM) FOR SOME FORECAST HORIZONS  $k$ .



(Solid line:  $k = 1$ , long dashes:  $k = 6$ , short dashes:  $k = 24$ , dots:  $k = 36$ , dashes and dots:  $k = 48$ ).

bivariate normal distribution with the estimated covariance matrix. Using this DGP with  $b = 0$ , the finite-sample distributions were first computed to obtain “size-adjusted” critical values. Although for the KPSS test such size adjustment is valid, this is not the case for the  $t$ -test. As discussed in section II, parametric bootstrap cannot be used to obtain the null distribution of the  $t$ -statistic in the local-to-unity case, and hence the size-adjusted critical values must be interpreted only as those of the infeasible  $t$ -test with the values of the nuisance parameters set at their estimated values. Therefore, the results cannot be used to compare the power of the KPSS test to that of the actual  $t$ -test but of the infeasible  $t$ -test with known local-to-unity parameter,  $c$ . For the sup- $c$  test, this kind of finite-sample size adjustment is, of course, not feasible. The results for some forecast horizons and 540 observations are presented in figure 2. For short forecast horizons, the infeasible  $t$ -test is very powerful, but the power begins to drop as the horizon exceeds 24 months. The power of the KPSS test, on the other hand, is good only for the short horizons. The fact that the power declines as the forecast horizon gets longer presumably follows from the problems caused by the increasing autocorrelation in the error term of the forecasting regression.

#### IV. Conclusion

In this paper, we have considered the problem of stock return predictability. It was pointed out that a potential explanation for the observed overrejection in regressions of the stock return on various forecasting variables is the high persistence of these forecasting variables, which causes the normal distribution to be a poor approximation of the distribution of the  $t$ -statistic under the null hypothesis of no predictability. Because in these cases the return is supposed

to be strictly stationary under the null hypothesis and a near unit root process plus noise under the alternative, we suggested testing for the stationarity of the stock return series instead. Because of the remarkable overlap in long-horizon regression, the well-known KPSS test seems to be the most convenient of the available stationarity tests, and it was employed in the empirical part of the paper. For comparison, the sup- $c$  test due to Cavanagh et al. (1995) was also used.

In the empirical analysis, monthly U.S. stock return series from the period 1928–1996 were considered. Not unexpectedly, predictability of these returns by the log dividend-price ratio using standard inference was established. The sup- $c$  test also indicated some predictability. Conversely, the stationarity tests did not find any predictability, implying that the stock returns could not be forecast by any of the typically considered highly persistent forecasting variables, including the dividend-price ratio. Small-sample simulation experiments confirmed that there indeed exists an overrejection problem with standard inference in these models and that the new approach controls size much better. The simulation results also lend support to the conjecture that, to some extent, the overrejection at longer forecast horizons is caused by error autocorrelation due to using overlapping observations. The local asymptotic power calculations show that, although the stationarity test is fairly powerful, the power loss stemming from not knowing the value of the local-to-unity parameter can be considerable, and this was also indicated by the small-sample simulation experiments.

In conclusion, the stationarity test indicates no predictability of stock returns, and it controls size asymptotically as well as in finite samples, but it may lack power. The standard  $t$ -test, on the other hand, suggests predictability, but it is not asymptotically valid and severely overrejects in finite samples, even when using conservative sup- $c$  critical values. Unfortunately, the power properties of the  $t$ -test cannot be evaluated using Monte Carlo simulation methods because it appears difficult to find proper size adjustment, and, therefore, this conflict cannot easily be resolved. Some support for the no-predictability result is provided by the fact that Wolf (2000) also did not find any predictability of U.S. stock returns by the dividend-price ratio using the subsampling method that was shown to be superior to the standard  $t$ -test once one is willing to exclude the potential exact unit root in the process of the dividend-price ratio.

#### REFERENCES

- Andrews, D. W. K., "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica* 59:3 (1991), 817–858.
- Basawa, I. V., A. K. Mallik, W. P. McCormick, J. H. Reeves, and R. L. Taylor, "Bootstrapping Unstable First-Order Autoregressive Processes," *Annals of Statistics* 19:2 (1991), 1098–1101.
- Campbell, J. Y., A. W. Lo, and A. C. MacKinlay, *The Econometrics of Financial Markets* (Princeton: Princeton University Press, 1987).
- Cavanagh, C. L., G. Elliott, and J. H. Stock, "Inference in Models with Nearly Integrated Regressors," *Econometric Theory* 11:5 (1995), 1131–1147.
- Choi, I., "Residual Based Tests for the Null of Stationarity with Applications to U.S. Macroeconomic Time Series," *Econometric Theory* 10:3/4 (1994), 720–746.
- Elliott, G., and J. H. Stock, "Inference in Time Series Regression when the Order of Integration of a Regressor is Unknown," *Econometric Theory* 10:3/4 (1994), 672–700.
- Goetzmann, W. N., and P. Jorion, "Testing the Predictive Power of Dividend Yields," *Journal of Finance* 48:2 (1993), 662–679.
- Hobijn, B., P. H. Franses, and M. Ooms, "Generalizations of the KPSS-Test for Stationarity," Econometric Institute Report no. 9802/A (1998).
- Hodrick, R. J., "Dividend Yields and Expected Stock Returns: Alternative Procedures for Inference and Measurement," *Review of Financial Studies* 5:3 (1992), 357–386.
- Kaul, G., "Predictable Components in Stock Returns" (pp. 269–296), in G. S. Maddala and C. R. Rao (Eds.), *Statistical Methods in Finance, Handbook of Statistics*, vol. 14 (Amsterdam: Elsevier, 1994).
- Kim, M. J., C. R. Nelson, and R. Startz, "Mean Reversion in Stock Prices? A Reappraisal of the Empirical Evidence," *Review of Economic Studies* 58:3 (1991), 515–528.
- Kirby, C., "Measuring the Predictable Variation in Stock and Bond Returns," *Review of Financial Studies* 10:3 (1997), 579–630.
- , "The Restrictions on Predictability Implied by Rational Asset Pricing Models," *Review of Financial Studies* 11:2 (1998), 343–382.
- Kwiatkowski, D., P. C. B. Phillips, P. Schmidt, and Y. Shin, "Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root: How Sure Are We that Economic Time Series Have a Unit Root?" *Journal of Econometrics* 54:1–3 (1992), 129–178.
- Mankiw, N. G., and M. D. Shapiro, "Do We Reject Too Often?" *Economics Letters* 20:2 (1986), 139–145.
- Nabeya, S., and B. E. Sørensen, "Asymptotic Distributions of the Least-Squares Estimators and Test Statistics in the Near Unit Root Model with Non-Zero Initial Value and Local Drift and Trend," *Econometric Theory* 10:5 (1994), 937–966.
- Nelson, C. R., and M. J. Kim, "Predictable Stock Returns: The Role of Small Sample Bias," *Journal of Finance* 48:2 (1993), 641–661.
- Newey, W. K., and K. D. West, "Automatic Lag Selection in Covariance Matrix Estimation," *Review of Economic Studies* 61:4 (1994), 631–653.
- Nyblom, J., and T. Mäkeläinen, "Comparisons of Tests for the Presence of Random Walk Coefficients in a Simple Linear Model," *Journal of the American Statistical Association* 78:384 (1983), 856–864.
- Phillips, P. C. B., "Toward a Unified Theory for Autoregression," *Biometrika* 74:3 (1987), 535–547.
- Richardson, M., and J. H. Stock, "Drawing Inferences from Statistics Based on Multiyear Asset Returns," *Journal of Financial Economics* 25:2 (1989), 323–348.
- Saikkonen, P., and R. Luukkonen, "Testing for Moving Average Unit Root in Autoregressive Integrated Moving Average Models," *Journal of the American Statistical Association* 88:422 (1993), 596–601.
- Stambaugh, R. F., "Bias in Regression with Lagged Stochastic Regressors," CRSP working paper no. 156 (1986).
- Stock, J. H., "Confidence Intervals for the Largest Autoregressive Root in U.S. Macroeconomic Time Series," *Journal of Monetary Economics* 28:3 (1991), 435–459.
- , "Cointegration, Long-Run Comovements, and Long-Horizon Forecasting" (pp. 34–60), in D. M. Kreps and Kenneth F. Wallis (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Seventh World Congress*, vol. III (Cambridge, UK: Cambridge University Press, 1997).
- Stock, J. H., and M. W. Watson, "Confidence Sets in Regressions with Highly Serially Correlated Regressors," Harvard University manuscript (1996).



Wolf, M., "Stock Returns and Dividend Yields Revisited: A New Way to Look at an Old Problem," *Journal of Business & Economic Statistics* 18:1 (2000), 18–30.

Wright, J. H., "Confidence Sets for Cointegrating Coefficients Based on Stationarity Tests," *Journal of Business & Economic Statistics* 18:2 (2000), 211–222.

APPENDIX

Consistency and Local Asymptotic Power of the  $L^\mu$  Test

This appendix is closely related to Appendix A of Wright (2000), in which similar results for cointegrated systems are derived. Assuming that the lag truncation parameter  $p$  in estimating the long-run variance of the error term  $u_{2t}$  is  $o_p(T)$ , the KPSS test is consistent against fixed alternatives  $\beta \neq 0$  in equation (1) when  $x_t$  is a near unit root process. This can be seen as follows. The Bartlett kernel estimator in the denominator of  $L^\mu$  is

$$\hat{\sigma}_{u_2}^2 = \hat{\gamma}_0 + 2 \sum_{j=1}^p \left(1 - \frac{j}{p+1}\right) \hat{\gamma}_j,$$

where  $\hat{\gamma}_j$  is the  $j$ th autocovariance. Because  $|\hat{\sigma}_{u_2}^2| \leq (2p+1)\hat{\gamma}_0$  and, for a near unit root process,  $\hat{\gamma}_0 = O_p(T)$ ,  $\hat{\sigma}_{u_2}^2 = O_p(pT)$ . The numerator of  $L^\mu$  is

$$\begin{aligned} & T^{-2} \sum_{t=1}^T \left( \sum_{s=1}^t (r_s^k - \bar{r}^k) \right)^2 \\ &= T^{-1} \sum_{t=1}^T \left( T^{-1/2} \sum_{s=1}^t (r_s^k - \bar{r}^k) \right)^2 \\ &= T^{-1} \sum_{t=1}^T \left( T^{-1/2} \sum_{s=1}^t u_{2s} + T^{-1/2} \beta \sum_{s=1}^t x_{s-k}^\mu \right)^2, \end{aligned}$$

where  $\mu$  refers to OLS demeaning. The latter term in the parentheses is  $O_p(T)$ , and hence  $L^\mu = O_p(T/p)$  so that it diverges to infinity, provided  $p = o(T)$ .

Let us next consider the asymptotic power of the KPSS test against local alternatives  $\beta = b/T$ . First, we show the consistency of the Bartlett kernel estimator of the long-run variance computed under the local alternative,  $\hat{\sigma}_{u_2}^2$ , assuming  $p = o(T)$ . Here, the  $j$ th sample autocovariance is

$$\begin{aligned} \tilde{\gamma}_j &= T^{-1} \sum_{t=j+1}^T (r_t^k - \bar{r}^k)(r_{t-j}^k - \bar{r}^k) \\ &= T^{-1} \sum_{t=j+1}^T (bT^{-1}x_{t-1}^\mu + u_{2t} - \bar{u}_2)(bT^{-1}x_{t-j-1}^\mu + u_{2t-j} - \bar{u}_2) \\ &= T^{-1} \sum_{t=j+1}^T (u_{2t} - \bar{u}_2)(u_{2,t-j} - \bar{u}_2) + O_p(T^{-1}) \end{aligned}$$

and

$$\hat{\sigma}_{u_2}^2 = \tilde{\gamma}_0 + 2 \sum_{j=1}^p (1 - j/(p+1))\tilde{\gamma}_j = \hat{\sigma}_{u_2}^2 + O_p(p/T)$$

which converges stochastically to  $\omega_{22}$  provided  $p = o(T)$ . The denominator of  $L^\mu$  is

$$\begin{aligned} & T^{-2} \sum_{t=1}^T \left( \sum_{s=1}^t (b(T^{-1}x_{s-1} - T^{-1}\bar{x}) + u_{2s} - \bar{u}_2) \right)^2 \\ &= T^{-1} \sum_{t=1}^T \left( b(T^{-3/2} \sum_{s=1}^t x_{s-1} - (t/T)T^{-3/2} \sum_{i=1}^T x_i \right. \\ &\quad \left. + T^{-1/2} \sum_{s=1}^t u_{2s} - (t/T)T^{-1/2} \sum_{i=1}^T u_{2i} \right)^2 \Rightarrow \\ &\quad \int_0^1 \left( b\omega_{11}^{1/2} \left( \int_0^r J_c(s) ds - r \int_0^1 J_c(s) ds \right) \right. \\ &\quad \left. + \omega_{22}^{1/2}(W_2(r) - rW_2(1)) \right)^2 dr. \end{aligned}$$

Combining these two results, we get the asymptotic distribution of  $L^\mu$  under local alternatives,

$$\begin{aligned} L^\mu &\Rightarrow \int_0^1 \left( b(\omega_{11}/\omega_{22})^{1/2} \left( \int_0^r J_c(s) ds - r \int_0^1 J_c(s) ds \right) \right. \\ &\quad \left. + W_2(r) - rW_2(1) \right)^2 dr. \end{aligned}$$

This indicates that local asymptotic power depends on the ratio of the long-run variances of the error terms of the equations for the forecasting variable and the return. Specifically, the higher  $\omega_{11}$  is in relation to  $\omega_{22}$ , the more powerful the test is.