ELSEVIER

Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis

Todd E. Clark[a],[*], Kenneth D. West[b]

[a]*Economic Research Department, Federal Reserve Bank of Kansas City, 925 Grand Blvd., Kansas City, MO 64198, USA*
[b]*Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison, WI 53706-1393, USA*

Abstract

We consider using out-of-sample mean squared prediction errors (MSPEs) to evaluate the null that a given series follows a zero mean martingale difference against the alternative that it is linearly predictable. Under the null of no predictability, the population MSPE of the null ''no change'' model equals that of the linear alternative. We show analytically and via simulations that despite this equality, the alternative model's sample MSPE is expected to be *greater* than the null's. For rolling regression estimators of the alternative model's parameters, we propose and evaluate an asymptotically normal test that properly accounts for the upward shift of the sample MSPE of the alternative model. Our simulations indicate that our proposed procedure works well.
© 2005 Elsevier B.V. All rights reserved.

---

*Corresponding author. Tel.: +1 816 881 2575; fax: +1 816 881 2199.
E-mail addresses: todd.e.clark@kc.frb.org (T.E. Clark), kdwest@wisc.edu (K.D. West).

1. Introduction

This paper considers the use of mean squared prediction errors (MSPEs) to evaluate the null that a given series follows a martingale difference against the alternative that the series is linearly predictable. A prominent application is to forecasts of changes in floating exchange rates. A long literature has used MSPEs to consider exchange rate prediction, generally finding that one cannot reject the null that exchange rate changes are unpredictable. Meese and Rogoff (1983) is the seminal reference on tests of what has come to be called the random walk model of the exchange rate; Cheung et al. (2003) is a recent update.

Our paper is applicable more generally to financial or other series whose changes are presumed to be unpredictable. For differences of such series, the null prediction at any horizon is a value of zero (that is, one predicts "no change" for the series whose difference is being predicted). Under the alternative, one uses a linear model to predict the differenced series. Under the null, the population regression coefficients in the alternative model are zero. This implies that under the null the population MSPEs of the two models are equal and the difference in MSPEs is zero.

Clive Granger's pioneering work on forecast evaluation proposed methods to formally evaluate the null that the difference in MSPEs from two such competing models is zero (Granger and Newbold, 1977; Ashley et al., 1980). A recent literature has suggested methods that extend Granger's work by applying results from generalized method of moments (Hansen, 1982). It appears that the first papers to so were Meese and Rogoff (1988), Christiano (1989) and West et al. (1993), with Diebold and Mariano (1995) and West (1996) offering general procedures for forecast evaluation. Among other advantages, the generalized method of moments approach allows for conditional heteroskedasticity in forecast errors.

Results in these papers are not, however, adequate for the applications that we consider. For in tests of the martingale difference hypothesis, as in many other applications, the models being compared are nested. West (1996, p. 1073) notes specifically that his procedures maintain a rank condition that is not satisfied when models are nested. It is our reading that a similar rank condition is implicit in Diebold and Mariano (1995). Simulations in McCracken (2004) and Clark and McCracken (2001, 2003) demonstrate that when comparing nested models, mechanical application of the procedures in West (1996) or Diebold and Mariano (1995) leads to very poorly sized tests; McCracken (2004) develops an asymptotic approximation that rationalizes the non-normal distribution of test statistics revealed by those simulations, while Chao et al. (2001) propose an alternative test that is asymptotically chi-squared under general conditions.

Our contribution is twofold. The first is to make a general observation about the center of the distribution of the difference in MSPEs when models are nested. For the most part, the previous literature has implicitly or explicitly assumed that since the difference in population MSPEs is zero, the difference in sample MSPEs should be approximately zero. We show that under the null, the sample MSPE from the alternative model is expected to be *greater* than that of the null. Intuitively, this is because the alternative model introduces noise into its forecasts by

estimating a parameter vector that in fact is not helpful in prediction.[1] The implication is that a finding that the null model has a smaller MSPE should *not* be taken as prima facie evidence against the alternative that the series in question is linearly predictable.

Our second contribution is to propose and evaluate a testing procedure that properly adjusts for the upward shift in the alternative model's MSPE. This procedure, which is applicable when forecasts are obtained from rolling regressions, is called "*MSPE-adjusted*." In our simple setup, the expected value of the upward shift is estimable in an obvious fashion. One therefore adjusts the MSPE from the alternative model downwards by a suitable estimate. One then compares MSPEs from the null and (adjusted) alternative. In this final comparison, inference may proceed in straightforward fashion using conventional asymptotically normal procedures familiar from Diebold and Mariano (1995) or West (1996). Alternatively, one may use bootstrapped critical values in the final comparison.

We evaluate our procedure, and compare it to some others, with simulations, focusing on conventional asymptotic inference. We find that our procedure is reasonably well-sized in samples of size typically available. Nominal 0.10 tests have actual size of around 0.07–0.09. This applies even for data that are plausibly calibrated and therefore messy: we allow the predictor in the alternative model to have a near unit root, as seems to be the case for an interest rate differential in our exchange rate application; we allow the martingale difference series to display conditional heteroskedasticity, as seems to be the case for high frequency exchange rate data.

Such simulation results distinguish our procedure from one we call "MSPE-normal." MSPE-normal looks at differences in MSPEs that are constructed without adjusting for the expected upward shift in the alternative model's MSPE, and are sometimes called Diebold and Mariano (1995) or DM statistics. MSPE-normal behaves as documented in McCracken (2004) and Clark and McCracken (2001, 2003): use of standard normal critical values generally results in very poorly sized tests, with nominal 0.10 tests generally having actual size less than 0.02. By contrast, and consistent with the results in Clark and McCracken (2001), use of the asymptotic values of the non-standard distribution tabulated in McCracken (2004) results in nicely sized tests. (At least with our data and sample sizes, then, the non-normality emphasized by McCracken (2004) and Clark and McCracken (2001, 2003) is substantially attributable to the noise in the alternative MSPE.) All tests seem to display size adjusted power that is roughly equivalent.

Some recent applied literature, which has absorbed the findings in McCracken (2004), has noted the inadvisability of using standard critical values when comparing nested models. One part of this literature then proceeds to use those critical values, albeit sometimes apologetically (e.g., Clarida et al., 2003; Cheung et al., 2003). Another part of this literature builds on Mark (1995) and Kilian (1999) to construct critical values with simulation or bootstrap techniques (e.g., Kilian and Taylor, 2003;

---

[1]This result is related but not identical with asymptotic results on MSPEs from overparameterized autoregressive models (Kunitomo and Yamamoto, 1985; Ing, 2003). See Section 2.

McCracken and Sapp, 2005). In our simulations, bootstrap procedures display size that is slightly better and power that is similar to our test with conventional asymptotic inference. Since the conventional asymptotic version is markedly more simple computationally, it is our belief that many will find it preferable to apology or bootstrapping.

We illustrate our methodology using monthly changes in four bilateral exchange rate series, the US dollar against the Canadian dollar, Japanese yen, British pound and Swiss franc. In accordance with interest parity, the alternative model uses an interest rate differential to predict exchange rate changes. The point estimate of the MSPE is smaller for the martingale difference model than for the interest parity model for three of the four series, the exception being the Canadian dollar. But upon adjusting for the upward shift in the alternative MSPE, we find that we can reject the null of the martingale difference model at the 5 percent level for not only the Canadian dollar but also for the Swiss franc. These rejections contrast with a long literature that finds little or no fault with the "random walk" model of the exchange rate.

To prevent confusion, we note that our results apply specifically to nested models. The analytical and simulation results in West (1996) suggest that when comparing MSPEs from non-nested models, one need not adjust as proposed here. One can instead construct a test statistic in familiar fashion. We also advise the reader that we have aimed for clarity at the expense of generality. For example, we assume stationarity and a linear alternative, even though generalizations are straightforward. As well, we caution the reader that we do not attempt to rationalize or motivate use of out-of-sample tests, nor to make recommendations for optimal choice of horizon, rolling regression sample size and so on. Our aim is to supply useful tools to the many researchers who find such tests of interest. On the tradeoff between in and out-of-sample tests, see Inoue and Kilian (2004); on selection of rolling regression sample size, see Clark and McCracken (2004). Finally, we suggest our procedure for an investigator testing the martingale difference against one or a handful of alternative models. For testing of all aspects of a martingale difference or general unpredictability hypothesis, see de Jong (1996) and Corradi and Swanson (2002).

Section 2 of the paper motivates our test. Section 3 details construction of our test statistic. Section 4 presents simulation results. Section 5 presents our empirical example. Section 6 concludes. An appendix available on the authors's websites presents detailed simulation results.

2. Mean squared prediction errors in nested models

We wish to evaluate the parsimonious possibility that a scalar variable y_t is a zero mean martingale difference. For such a model, one simply forecasts the future value of y_t to be zero. That forecast is optimal for any horizon and for any past data on y_t and related variables. An alternative model posits that y_t is linearly related to a vector of variables X_t. Thus the null model (model 1) and the alternative

model (model 2) are:

$$y_t = e_t \text{ (model 1: null model),} \tag{2.1}$$

$$y_t = X_t'\beta + e_t \text{ (model 2: alternative model).} \tag{2.2}$$

Under the null, $\beta = 0$; under the alternative, $\beta \neq 0$. Let E_{t-1} denote expectations conditional on current and past X's and past e's: $E_{t-1}e_t \equiv E(e_t| X_t, e_{t-1}, X_{t-1}, e_{t-2}, \ldots)$. We assume that under both the null and the alternative, e_t is a zero mean martingale difference:

$$E_{t-1}e_t \equiv E(e_t|X_t, e_{t-1}, X_{t-1}, e_{t-2}, \ldots) = 0. \tag{2.3}$$

Since e_t has conditional mean zero, it is serially uncorrelated. It may, however, be conditionally heteroskedastic, with $E_{t-1}e_t^2 \neq Ee_t^2$. Our dating allows (indeed, presumes) that X_t is observed prior to y_t. For example, the alternative model may be an AR(1) for y_t, in which case X_t is 2×1 with $X_t' = (1, y_{t-1})$.

We assume that one evaluates the null via comparison of out-of-sample mean squared prediction errors (MSPEs). (For the remainder of the paper, "MSPE" without qualification refers to an out-of-sample statistic.) For simplicity, we focus on one step ahead predictions, briefly discussing multistep predictions when we turn to inference. One has a sample of size $T + 1$. The last P observations are used for predictions. The first prediction is for observation $R + 1$, the next for $R + 2, \ldots$, the final for $T + 1$. We have $T + 1 = R + P$. We take P and R as given. For $t = R$, $R + 1, \ldots, T$, one uses data prior to t to predict y_{t+1}. Let $\hat{\beta}_t$ denote a regression estimate of β that relies on data prior to t, putting aside the details of the sample used to construct $\hat{\beta}_t$ (rolling vs. recursive).

The one step ahead prediction for model 1 is a constant value of 0, for model 2 is $X_{t+1}'\hat{\beta}_t$. The corresponding prediction errors are y_{t+1} and $y_{t+1} - X_{t+1}'\hat{\beta}_t$, with MSPEs

$$\hat{\sigma}_1^2 \equiv P^{-1}\Sigma_{t=T-P+1}^{T} y_{t+1}^2 = \text{MSPE from model 1,} \tag{2.4}$$

$$\hat{\sigma}_2^2 \equiv P^{-1}\Sigma_{t=T-P+1}^{T} (y_{t+1} - X_{t+1}'\hat{\beta}_t)^2 = \text{MSPE from model 2.} \tag{2.5}$$

Under the null, $\beta = 0$ and population MSPEs for the two models are equal: $Ey_{t+1}^2 - E(y_{t+1} - X_{t+1}'\beta)^2 = 0$. We wish to use $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ to test the null. We turn to asymptotic theory to understand the large sample ($T \to \infty$) behavior of $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$, the difference in MSPEs.

One possibility is to apply the procedures proposed by Ashley et al. (1980), Diebold and Mariano (1995) or West (1996). Observe that under the null, $\beta = 0$, implying that in population the MSPEs from the two models are equal. Observe as well that $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$ is constructed by averaging P observations on prediction errors. One might interpret these papers as suggesting that one treat $\sqrt{P}(\hat{\sigma}_1^2 - \hat{\sigma}_2^2)$ as asymptotically normal with a variance–covariance matrix that can be estimated in familiar fashion.[2]

---

[2]Some applied papers such as Goyal and Welch (2003) have interpreted Diebold and Mariano in just this way. West (1996, p. 1073) explicitly states that his asymptotic results presume the two models are non-nested. So the technical conditions in West (1996) exclude applications such as ours, because our models are nested.

Some simple algebra suggests that such an interpretation is questionable. With straightforward algebra, the difference in MSPEs is

$$\hat{\sigma}_1^2 - \hat{\sigma}_2^2 = 2\left(P^{-1} \sum_{t=T-P+1}^{T} y_{t+1} X_{t+1}' \hat{\beta}_t\right) - \left[P^{-1} \sum_{t=T-P+1}^{T} (X_{t+1}' \hat{\beta}_t)^2\right]. \qquad (2.6)$$

Under the null, $y_{t+1} = e_{t+1}$ and $Ee_{t+1} X_{t+1}' \hat{\beta}_t = 0$. So under the null we expect $2(P^{-1}\sum_{t=T-P+1}^{T} y_{t+1} X_{t+1}' \hat{\beta}_t) \approx 0$. Since $-P^{-1}\sum_{t=T-P+1}^{T} (X_{t+1}' \hat{\beta}_t)^2 < 0$ by construction, we expect to find $\hat{\sigma}_1^2 < \hat{\sigma}_2^2$—that is, *under the null we expect the MSPE of the null model to be smaller than that of the alternative model*. This follows because in finite samples, the alternative model's MSPE is expected to be pushed upwards by the noise term $P^{-1}\sum_{t=T-P+1}^{T} (X_{t+1}' \hat{\beta}_t)^2$. We emphasize that this result applies regardless of whether rolling, recursive or fixed schemes are used to obtain $\hat{\beta}_t$.

Extensive simulations in McCracken (2004) indicate that for nested models such as ours the distribution suggested by Diebold and Mariano (1995) provides a poor approximation to the actual finite sample distribution of $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$. In particular, the mean and median of $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$ are not zero, as suggested by that asymptotic approximation, but instead are negative, as suggested by the simple algebra above. That is, on average, across simulations, model 1's MSPE is less than model 2's MSPE. In McCracken's simulations, the downward shift in the distribution of $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$ is more pronounced the larger is the number of regressors in model 2 (the larger is the dimension of X_t).

McCracken (2004) rationalizes these features with an asymptotic theory that assumes that the number of predictions is large ($P \to \infty$) and that the size of the samples used to estimate the sequence of $\hat{\beta}_t$'s is large ($R \to \infty$), with P/R approaching a finite non-zero constant. (Note that under this approximation, $\hat{\beta}_t \to_p 0$ so $\hat{\sigma}_1^2 - \hat{\sigma}_2^2 \to_p 0$). Unlike West (1996), McCracken (2004) assumes nested rather than non-nested models. He writes the limiting distribution as functionals of Brownian motion. Several papers have used similar approximations. The most relevant for our topic are the papers by Clark and McCracken (2001, 2003). Clark and McCracken verify via simulation that in comparison of nested models, application of the procedures of Diebold and Mariano (1995) or West (1996) leads to distinct departures from the asymptotic approximations those papers suggest.

We, too, will use asymptotic theory to capture such simulation results. To do so, we now restrict the investigator to have obtained the sequence of $\hat{\beta}_t$'s with rolling regression. (See points 4 and 5 below on implications of use of other schemes to obtain the $\hat{\beta}_t$'s.) Our asymptotic approximation follows Giacomini and White (2003) in holding R, the size of the regression sample, fixed, while letting the number of predictions P go to infinity. Thus, $R/P \to 0$, $R/T \to 0$. We propose an asymptotic thought experiment with R fixed because, as we shall see when we discuss simulation results, the implied asymptotic distribution well-approximates finite sample results for our test.

The alternative model's parameter vector $\hat{\beta}_t$ is estimated from rolling regressions that rely on data from periods $t - R + 1$ to t,

$$\hat{\beta}_t = \left(\sum_{s=t-R+1}^{t} X_s X_s' \right)^{-1} \sum_{s=t-R+1}^{t} X_s y_s.$$

(Note that, following conventions in the literature, R does double duty: observations 1 to R constitute the first regression sample, and R is the size of subsequent regression samples as well.) When R is fixed, $\hat{\beta}_t$ and $X_{t+1}'\hat{\beta}_t$ are well-behaved random variables under mild conditions. For expositional ease, we make the strong assumption that y_{t+1} and $X_{t+1}'\hat{\beta}_t$ are covariance stationary. Such an assumption perhaps is unappealing, since heterogeneity over time is a possible motivation for use of rolling samples. But while we have not checked all the details, it appears that we could accommodate heterogeneity such as that in Giacomini and White (2003). Giacomini and White (2003) analyze what they call "conditional" tests, a class of tests that excludes the one we study.) Such heterogeneity involves controlled variation in moments. The work of Inoue and Kilian (2004) indicates that sudden one-time jumps are not easily accommodated. Under stationarity, we will have

$$P^{-1} \sum_{t=T-P+1}^{T} (X_{t+1}'\hat{\beta}_t)^2 \to_p \mathrm{E}(X_{t+1}'\hat{\beta}_t)^2,$$

$$P^{-1} \sum_{t=T-P+1}^{T} e_{t+1} X_{t+1}'\hat{\beta}_t \to_p 0,$$

$$\hat{\sigma}_1^2 \to_p \sigma_1^2 \equiv \mathrm{E}y_t^2,$$

$$\hat{\sigma}_2^2 \to_p \mathrm{E}(y_{t+1} - X_{t+1}'\hat{\beta}_t)^2 \equiv \hat{\sigma}_2^2(R). \tag{2.7}$$

In (2.7), the "R" in $\hat{\sigma}_2^2(R)$ emphasizes that the large sample limit of $\hat{\sigma}_2^2$ depends on the regression sample size. For very large R, the sequence of $\hat{\beta}_t$'s will generally be very small, and $\hat{\sigma}_2^2(R)$ will be close to σ_1^2. But for small as well as large values of R, it follows from (2.6) and (2.7) that

$$\hat{\sigma}_1^2 - \hat{\sigma}_2^2 \to_p - \mathrm{E}(X_{t+1}'\hat{\beta}_t)^2 < 0. \tag{2.8}$$

It follows that we will tend to see $\hat{\sigma}_1^2 < \hat{\sigma}_2^2$, at least when the number of predictions P is large. Thus our approximation, which reflects the familiar dictum that a correctly parameterized model tends to outpredict an over-parameterized model, is capable of rationalizing the simulation evidence in McCracken (2004) and Clark and McCracken (2001, 2003) that, on average, $\hat{\sigma}_1^2 < \hat{\sigma}_2^2$.

Equivalently, let us rewrite (2.8) as

$$\sigma_1^2 = \sigma_2^2(R) - \mathrm{E}(X_{t+1}'\hat{\beta}_t)^2, \text{ or plim } \hat{\sigma}_1^2 - \mathrm{plim}[\hat{\sigma}_2^2 - \mathrm{E}(X_{t+1}'\hat{\beta}_t)^2] = 0.$$

We can then state: asymptotically, the MSPE of model 2 (the over-parameterized model) exceeds the MSPE of the correctly parameterized model by a quantity that reflects spurious small-sample fit.

To illustrate these points graphically, let us present smoothed density plots of some simulation results. Details on the data generating process (labeled "DGP 1" in the header to the figures) are given in our discussion of Monte Carlo results below. Briefly, $y_t \sim$ i.i.d $N(0,1)$, X_t is 2×1 with $X_t = (1, x_{t-1})'$, $x_t \sim$ AR(1) with parameter 0.95, x and y independent. The persistence in x_t roughly matches our empirical application. In this application, y_t is the monthly change in a dollar exchange rate, x_t is monthly interest differentials. Our simulations vary the regression sample size R and the number of predictions P. It may help to note that in our empirical application, which involves four exchange rate series, the number of monthly observation is $R = 120$, while $P = 168$ for two exchange rates and $P = 228$ for the other two exchange rates.

Fig. 1 presents plots for $R = 120$, various values of P. Panel A presents plots of $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$. There is substantial probability that $\hat{\sigma}_1^2 - \hat{\sigma}_2^2 < 0$. For $P = 144$, for example, $\hat{\sigma}_1^2 - \hat{\sigma}_2^2 < 0$ in 89.8% of the simulations. As P gets bigger, the density of $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$ begins to narrow, with an increasing probability of $\hat{\sigma}_1^2 - \hat{\sigma}_2^2 < 0$. Fig. 1B plots the density of the average value of $(X'_{t+1}\hat{\beta}_t)^2$, i.e., of $P^{-1}\Sigma_{t=T-P+1}^{T}(X'_{t+1}\hat{\beta}_t)^2$, again for $R = 120$ and various values of P. (To make Fig. 1B readable, the scale is different from that of Fig. 1A.) In Fig. 1B, the distribution is skewed to the right, especially for small values of P. Fig. 1C subtracts Fig. 1B from Fig. 1A and plots the result on the same scale as Fig. 1A. That is, Fig. 1C plots $\hat{\sigma}_1^2 - [\hat{\sigma}_2^2 - P^{-1}\Sigma_{t=T-P+1}^{T}(X'_{t+1}\hat{\beta}_t)^2]$, which is the difference in MSPEs, adjusted for the estimate of $E(X'_{t+1}\hat{\beta}_t)^2$. In a table below, we write this more concisely as

$$\hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - \text{adj.}), \quad \hat{\sigma}_2^2 - \text{adj.} \equiv \hat{\sigma}_2^2 - P^{-1}\Sigma_{t=T-P+1}^{T}(X'_{t+1}\hat{\beta}_t)^2. \tag{2.9}$$

In (2.9), "−adj." stands for the *adjustment* made to the MSPE for model 2 and $\hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - \text{adj.})$ is the adjusted difference in MSPEs.

The adjusted difference plotted in Fig. 1C inherits a bit of the skewness of the estimate of $E(X'_{t+1}\hat{\beta}_t)^2$. But the median is closer to zero in panel C than in panel A. As well, Fig. 1C may be a bit misleading, since it may wrongly suggest a non-zero mean. In fact, the mean value of the simulation data matches our asymptotic theory virtually spot on. For all values of P, the mean value of $\hat{\sigma}_1^2 - [\hat{\sigma}_2^2 - P^{-1}\Sigma_{t=T-P+1}^{T}(X'_{t+1}\hat{\beta}_t)^2]$ is less than 0.001 in absolute value. Thus the fact that the median is less than zero is balanced by the skew in the right tail.

Fig. 2 present similar results when $P = 144$ and R varies. Consistent with comments above, panel A indicates that as R increases, the difference in MSPEs shifts towards zero; panel B that the mean of $(X'_{t+1}\hat{\beta}_t)^2$ falls. The difference in MSPEs adjusted for the estimate of $E(X'_{t+1}\hat{\beta}_t)^2$ again has a median less than zero, and is skewed right. But once again the mean is essentially zero, as predicted by our asymptotic approximation.
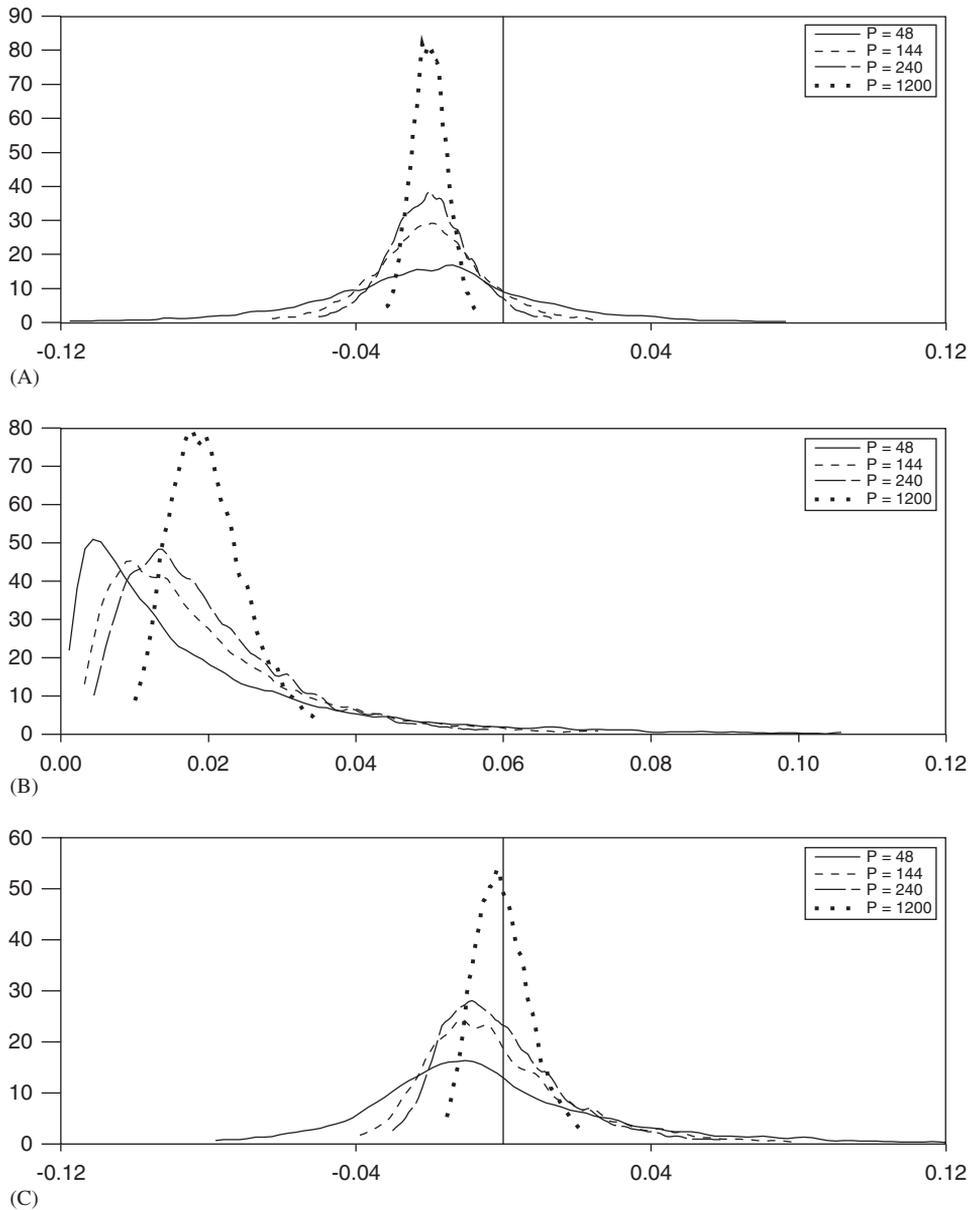
Fig. 1. Density of simulation MSPEs under the null, $R = 120$, P varying, DGP 1: (A) $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$; (B) Average of $(X'_{t+1}\hat{\beta}_t)^2$; (C) $\hat{\sigma}_1^2 - [\hat{\sigma}_2^2 - \text{average of } (X'_{t+1}\hat{\beta}_t)^2]$.

A central argument of this paper is that in testing the martingale difference hypothesis, one should not just compare MSPEs but instead adjust the comparison using an estimate of $E(X'_{t+1}\hat{\beta}_t)^2$. The practical importance of this proposal turns in
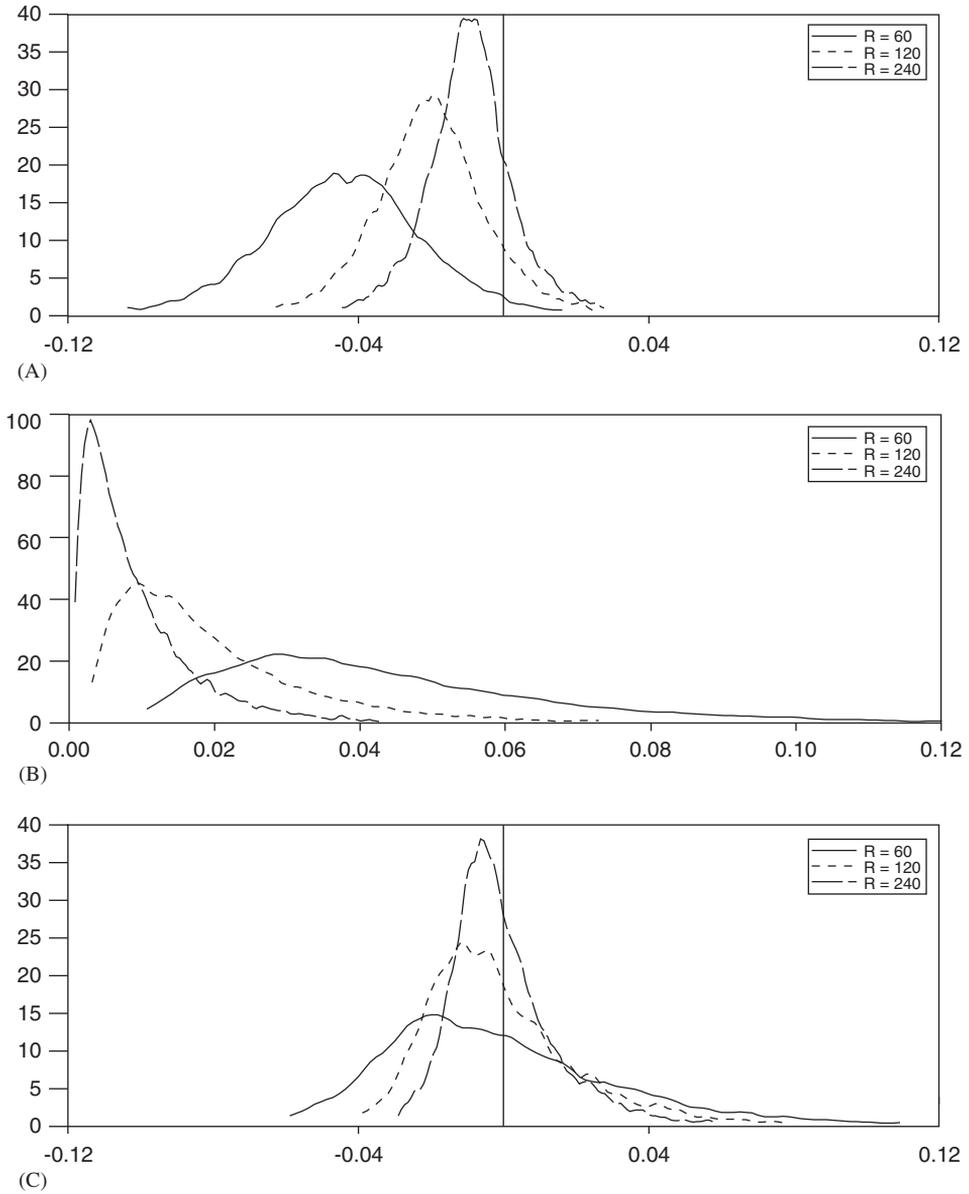
Fig. 2. Density of simulation MSPEs under the null, R varying, $P = 144$, DGP 1: (A) $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$; (B) Average of $(X'_{t+1}\hat{\beta}_t)^{\wedge 2}$; (C) $\hat{\sigma}_1^2 - [\hat{\sigma}_2^2 - \text{average of } (X'_{t+1}\hat{\beta}_t)^2]$.

part on the magnitude of $\mathrm{E}(X'_{t+1}\hat{\beta}_t)^2$ relative to σ_1^2. In light of our limited knowledge of the exact small sample behavior of least squares estimate $\hat{\beta}_t$ with time series data, it is hard to make analytical statements about the magnitude of $\mathrm{E}(X'_{t+1}\hat{\beta}_t)^2$ (equivalently, about $\sigma_2^2(R)$), and more generally the small sample behavior of

$\hat{\sigma}_1^2 - \hat{\sigma}_2^2$. But the following points may be helpful, and are consistent with the simulation evidence depicted in Figs. 1 and 2.

1. Consider first the uninteresting but tractable case in which X_{t+1} is univariate and consists solely of a constant. It is straightforward to show that in this case $-\mathrm{E}(X'_{t+1}\hat{\beta}_t)^2 = -\mathrm{E}y_t^2/R$. Recall that $\hat{\sigma}_1^2 \to_p \mathrm{E}y_t^2$. So with $R = 10$, we expect $\hat{\sigma}_1^2$, the null model's MSPE, to be about 10% smaller than $\hat{\sigma}_2^2$, the MSPE of a model that uses the sample mean (i.e., regresses on $X_{t+1} \equiv 1$) as the forecast. With $R = 100$ we expect $\hat{\sigma}_1^2$ to be about 1% smaller than $\hat{\sigma}_2^2$.

2. Consider now a case in which X_{t+1} includes a constant and $k - 1$ stochastic regressors. Let y_t and X_{t+1} be jointly normal, with y_t independent of X_{t+1} (in accordance with the null). To get a feel for the magnitude of $\mathrm{E}(X'_{t+1}\hat{\beta}_t)^2$ relative to σ_1^2 let us rewrite (2.8) as $1 - (\hat{\sigma}_2^2/\hat{\sigma}_1^2) \to_p - [\mathrm{E}(X'_{t+1}\hat{\beta}_t)^2/\sigma_1^2]$. Manipulation of textbook formulas for the normal regression model indicates that $\mathrm{E}(X'_{t+1}\hat{\beta}_t)^2/\sigma_1^2$ increases with the number of regressors k, and falls with the regression sample size R.[3] Our approximation thus accounts for McCracken's (2004) simulation evidence that $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$ lies farther below zero when X_t includes more regressors, a result that we have verified with our own set of simulations (details omitted). That $\mathrm{E}(X'_{t+1}\hat{\beta}_t)^2/\mathrm{E}y_t^2$ falls as R increases can be seen in Fig. 1C.

3. Some applied papers (see the introduction) have conducted hypothesis tests assuming $\sqrt{P}(\hat{\sigma}_1^2 - \hat{\sigma}_2^2)$ is asymptotically $N(0, W)$ with a covariance matrix W that can be estimated in standard fashion. Evidently, under our approximation, in which the asymptotic mean of $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$ is negative rather than zero (see (2.8)), $\sqrt{P}(\hat{\sigma}_1^2 - \hat{\sigma}_2^2)$ shifts increasingly below zero as P increases. When P is large, then, such tests will likely be especially poorly sized, a result confirmed by our simulations.

4. We have assumed that the sequence of $\hat{\beta}_t$'s is obtained with rolling regressions, partly to maintain contact with the relevant class of applications, partly for technical reasons. Some relevant applications obtain the sequence of $\hat{\beta}_t$'s not from rolling but recursive regressions. In such regressions, the sample size used to estimate the sequence of β's grows. The first sample uses data from 1 to R, the next from 1 to $R+1, \ldots,$ the final from 1 to T. The key expression (2.6) still holds when the $\hat{\beta}_t$'s are obtained in this fashion. And we will still have $\mathrm{E}e_{t+1}X'_{t+1}\hat{\beta}_t = 0$, $-[P^{-1}\Sigma_{t=T-P+1}^{T}(X'_{t+1}\hat{\beta}_t)^2] < 0$. Thus we have shown that with recursive estimation of β, we expect $\hat{\sigma}_1^2 < \hat{\sigma}_2^2$. But as a formal matter, we have not established the limiting distribution of suitably normalized $P^{-1}\Sigma_{t=T-P+1}^{T}y_{t+1}X'_{t+1}\hat{\beta}_t$; even in very simple cases, this distribution seems to be non-normal. See Clark and McCracken (2001, 2003).

---

[3] This statement relies on an approximation to the expectation of the multiple correlation coefficient in a least squares regression in which y and X are jointly normal, and y is independent of X (Theil, 1971, p. 191).

5. As well, some relevant applications predict with a single estimate, call it $\hat{\beta}_R$, rather than with a sequence of $\hat{\beta}_t$'s (e.g., Ashley et al., 1980). Forecast errors from the alternative model are then constructed as $y_{t+1} - X'_{t+1}\hat{\beta}_R$, $t = R, \ldots, R + P$; note that $\hat{\beta}_R$ maintains a fixed "R" subscript as t varies. Our results extend directly to such environments.

6. Using a high order approximation, and maintaining (in our notation) $R \rightarrow \infty$ and $P = 1$, Kunitomo and Yamamoto (1985) and Ing (2003) establish that when comparing a correctly specified with an overparameterized autoregressive model, the expected difference in MSPEs is negative. These papers do not discuss inference. On the other hand, these papers do allow the null model to make predictions with an estimated parameter vector, while we focus on a simple martingale difference as the null. We note that the simulations in McCracken (2004), which allow the null model to make predictions with an estimated parameter vector, find that the MSPE of the null model tends to be less than that of the alternative. Thus both analytical and simulation results suggest to us that our basic result generalizes to environments in which the null model relies on an estimated parameter vector.

7. Nonlinear models with additive errors are easily accommodated. Write the alternative model as $y_t = g_t(\beta) + e_t$, where the t subscript on g_t indicates dependence on data from period $t - 1$ and earlier. Possibilities for g_t include ARMA or Markov-switching models, and possible estimators include non-linear least squares, maximum likelihood and non-parametric estimators. In such cases, one simply replaces $X'_t\hat{\beta}_{t-1}$ in the formulas above with the period t fitted value of $g_t(\beta)$.

8. Let us summarize some central characteristics of our approach compared to that of McCracken (2004) and Giacomini and White (2003). We are unique in proposing that the investigator adjust the point estimate of the MSPE difference prior to performing inference. Like McCracken (2004), we examine the standard version of MSPE; Giacomini and White (2003) examine a class of tests that includes what Giacomini and White (2003) call a "conditional" MSPE but excludes the standard version of MSPE. Otherwise, we share central characteristics of Giacomini and White (2003) but not McCracken: like Giacomini and White (2003), we restrict ourselves to rolling schemes to obtain $\hat{\beta}_t$, but permit multiple as well as single period predictions, allow conditional heteroskedasticity, and accommodate a variety of estimators of $\hat{\beta}_t$; McCracken (2004) allows rolling and recursive schemes to obtain $\hat{\beta}_t$ but rules out multiperiod prediction horizons, conditional heteroskedasticity and any estimator but one with quadratic loss.[4] And at a technical level, we, like Giacomini and White (2003), use an asymptotic approximation that holds the regression sample size R fixed as $T \rightarrow \infty$, while McCracken assumes $R \rightarrow \infty$.

---

[4]Actually, McCracken (2004) allows non-quadratic loss in estimation of β as long as the loss function used in estimation is also used to evaluate forecasts. In our view, the practical implication is as stated in the text. Also, Clark and McCracken (2003) extend McCracken (2004) to allow multiperiod horizons.

3. The test statistic

We propose adjusting the difference in MSPEs by a consistent estimate of the asymptotic difference between the two. Hence we call our statistic "*MSPE-adjusted*." Thus we look at

$$\hat{\sigma}_1^2 - \left\{ \hat{\sigma}_2^2 - \left[P^{-1} \Sigma_{t=T-P+1}^{T} (X'_{t+1}\hat{\beta}_t)^2 \right] \right\} \equiv \hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - \text{adj.})$$
$$= (\text{say})\bar{f} \quad (\text{MSPE-adjusted}). \quad (3.1)$$

Then it follows from (2.6) that under mild conditions,

$$\sqrt{P}\bar{f} \sim_A N(0, V), \quad V = 4E(y_{t+1} X'_{t+1}\hat{\beta}_t)^2. \quad (3.2)$$

(As noted above, we assume stationarity for presentational ease.) One can use the obvious estimator of V, namely, $4P^{-1}\Sigma_{t=T-P+1}^{T}(y_{t+1} X'_{t+1}\hat{\beta}_t)^2$. One convenient way to construct both the point estimate and the test statistic is to define

$$\hat{f}_{t+1} \equiv y_{t+1}^2 - [(y_{t+1} - X'_{t+1}\hat{\beta}_t)^2 - (X'_{t+1}\hat{\beta}_t)^2]. \quad (3.3)$$

Then one can compute

$$\bar{f} = P^{-1}\Sigma_{t=T-P+1}^{T}\hat{f}_{t+1}, \quad \hat{V} = P^{-1}\Sigma_{t=T-P+1}^{T}(\hat{f}_{t+1} - \bar{f})^2. \quad (3.4)$$

We note that since $\bar{f} = 2(P^{-1}\Sigma_{t=T-P+1}^{T}y_{t+1}X'_{t+1}\hat{\beta}_t)$, our adjustment to the MSPE shows that a question about MSPEs is isomorphic to one about encompassing. Studies that have evaluated a statistic like $\bar{f}$ include Harvey et al. (1998), West and McCracken (1998), West (2001) and Clark and McCracken (2001, 2003). As well, our test is similar to that of Chao et al. (2001). When applied to our setting, Chao et al. (2001) propose examining the vector $P^{-1}\Sigma_{t=T-P+1}^{T}y_{t+1}X_{t+1}$. So the Chao et al. (2001) statistic, like $\bar{f}$, looks to correlations between y and X as a measure of the null. We examine a particular linear combination of the elements of X, because this linear combination lets us quantify the expected discrepancy between MSPEs. (Chao et al.'s (2001) technical conditions include $R \to \infty$ as $P \to \infty$ so under their conditions the asymptotic discrepancy between MSPEs is zero.)

For longer horizon forecasts, one can proceed as follows. The relevant class of applications typically uses overlapping data to make a τ-step ahead forecast of $y_{t+\tau} + y_{t+\tau-1} + \cdots y_{t+1} = (\text{say})\ y_{t+\tau,\tau}$. (In this notation, $y_{t+1,1} \equiv y_{t+1}$, and for $\tau = 1$ the formulas about to be presented reduce to those given above.) The MSPE from the null model is $(P - \tau + 1)^{-1}\Sigma_{t=T-P+1}^{T-\tau+1}y_{t+\tau,\tau}^2$. The alternative model regresses $y_{t+\tau,\tau}$ on a vector of regressors X_{t+1}. (Of course in some applications the vector of predictors in the alternative model will vary with the forecast horizon τ, and might properly be denoted $X_{t+1,\tau}$. We use the notation X_{t+1} to limit notational clutter.) The MSPE from the alternative is $(P - \tau + 1)^{-1}\Sigma_{t=T-P+1}^{T-\tau+1}(y_{t+\tau,\tau} - X'_{t+1}\hat{\beta}_t)^2$. One adjusts the difference in MSPEs exactly as in (3.1) above. With overlapping data, however, the resulting time series $(2y_{t+\tau,\tau}X'_{t+1}\hat{\beta}_t) \equiv [2(y_{t+\tau} + y_{t+\tau-1} + \cdots y_{t+1})X'_{t+1}\hat{\beta}_t]$, $t = T - P + 1, \ldots, T - \tau + 1$, follows a MA($\tau - 1$) process under the null. For $\tau > 1$, one needs to account for this serial correlation when performing inference.

One can use any one of a variety of popular estimators. We suggest a mean-adjusted version of the estimator developed in West (1997), which in this case reduces to a mean adjusted version of Hodrick (1992): Let $\hat{g}_t = 2y_t(X'_t\hat{\beta}_{t-1} + \cdots + X'_{t-\tau+1}\hat{\beta}_{t-\tau})$, with sample mean $\bar{g}$. Let

$$\hat{V} = (P - 2\tau + 2)^{-1}\Sigma_{t=T-P+\tau}^{T-\tau+1}(\hat{g}_{t+\tau} - \bar{g})^2. \tag{3.5}$$

Then $\hat{V}$ is a consistent estimator of the asymptotic variance of the adjusted difference in MSPEs.

4. Simulation results

We use Monte Carlo simulations of simple bivariate data-generating processes to evaluate the finite-sample size and power of our approach and compare it to some existing tests. The experiments incorporate features common in applications for which our suggested approach to inference may be most useful-exchange rate and other asset price applications in which the null hypothesis is that the predictand is a martingale difference. In such applications, the variance of the predictand y_t is very high relative to the variance of the predictors in X_t, and those predictors are highly persistent. We use DGPs with these features to evaluate several tests: our *MSPE-adjusted* test, an unadjusted MSPE test with standard normal critical values (called "*MSPE-normal*" in the tables, but commonly referred to as a Diebold–Mariano test), an unadjusted MSPE test with McCracken's (2004) non-standard critical values (*MSPE-McCracken* in the tables), and the Chao et al. (2001) test (CCS, in the tables).

4.1. Experimental design

The baseline DGPs share the same basic form, widely used in studies of the properties of predictive regressions (see, for example, Mankiw and Shapiro, 1986; Nelson and Kim, 1993; Stambaugh, 1999; Campbell, 2001; Tauchen, 2001):

$$y_t = bx_{t-1} + e_t, \quad X_t = (1, x_{t-1})', \quad x_t = 0.95x_{t-1} + v_t,$$
$$\mathrm{E}_{t-1}e_t = 0, \quad \mathrm{E}_{t-1}v_t = 0, \quad \mathrm{var}(e_t) = 1, \quad \mathrm{var}(v_t) = \sigma_v^2, \quad \mathrm{corr}(e_t, v_t) = \rho; \quad (4.1)$$

$$b = 0 \text{ in experiments evaluating size}, \quad b \neq 0 \text{ in experiments evaluating power}.$$

The null forecast (model 1) is simply the martingale difference or "no change" forecast of 0 for all t. The alternative forecast (model 2) is obtained from a regression of y_t on X_t, using a rolling window of R observations. (Because the alternative regression includes a constant, adding a constant to the x_t equation would have no effect on the results.)

In analyzing both size and power, we report results for two general parameterizations of the above DGP. The first parameterization, labeled DGP 1, is based roughly on estimates from the exchange rate application considered in the

empirical work reported in the next section: $\sigma_v = 0.025$ and $\rho = 0$.[5] The second parameterization, DGP 2, is calibrated to monthly excess returns in the S&P500 and the dividend–price ratio: $\sigma_v = 0.036$ and $\rho = -0.9$. For both DGPs, the coefficient b is set to 0 in the size experiments. In evaluating power, we consider values of b based roughly on estimates from the associated exchange rate and stock market data. Specifically, we simulate DGP 1 with $b = -2$ and DGP 2 with $b = 0.365$.

While we focus on results for data generated from homoskedastic draws from the normal distribution, we extend DGP 2 to consider data with conditional heteroskedasticity and fat tails-features that are often thought to characterize financial data. Select results are reported for experiments in which $e_t (= y_t)$ follows a GARCH(1, 1) process, parameterized according to estimates for excess returns in the S&P500:

$$y_t = e_t = \sqrt{h_t}\varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d. } N(0,1), \quad h_t = 0.05 + 0.85 h_{t-1} + 0.10 e_{t-1}^2,$$
$$\rho = -0.9. \tag{4.2}$$

With this parameterization, the unconditional variance of y is the same as in the homoskedastic case (equal to 1). Select results are also reported for experiments in which there is conditional heteroskedasticity in y, of a simple multiplicative form, scaled to keep the unconditional variance at unity:

$$y_t = e_t = \sqrt{h_t}\varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d. } N(0,1), \quad h_t = x_{t-1}^2/\text{var}(x_t), \quad \text{corr}(\varepsilon_t, v_t) = -0.9, \tag{4.3}$$

where $\text{var}(x_t) = \sigma_v^2/(1 - 0.95^2)$. Finally, in results not reported in the interest of brevity, we also consider data generated from a fat tailed i.i.d. distribution similar to one used in simulations in Diebold and Mariano (1995), in which the forecast error e_t follows a $t(6)$ distribution and the error v_t is a linear combination of underlying innovations drawn from the $t(6)$ distribution. The results for fat-tailed data prove to be very similar to those for normally distributed data.

To match the variety of settings that appear in empirical research, we consider a range of R and P values, with emphasis on P large relative to R. With monthly data in mind, R is set to 60, 120, or 240, and P is set to 48, 96, 144, 250, 480, and 1200.[6] For the given setting of R, a total of $R + 1200$ observations are generated. The initial observation on x is generated from its unconditional normal distribution. Rolling one step ahead predictions are formed for observations $R + 1$ through $R + 1200$, using a window of R observations. For each value of P, one step ahead predictions are evaluated from $R + 1$ through $R + P$. For multistep predictions of horizon τ, predictions are evaluated from $R + \tau$ through $R + P$, with the total number of predictions being $P - \tau + 1$. The number of simulations is 10,000.

---

[5]In the data, the correlation in the residuals tends to be small, although it varies in sign across countries. Simulations of a DGP with the same form but a small, non-zero ρ yielded very similar results.

[6]The values of P are chosen judiciously to facilitate the use of McCracken's (2004) critical values in evaluating the MSPE test. McCracken reports critical values for select values of $\pi \equiv \lim_{P,R \to \infty}(P/R)$.

4.2. Procedures for inference

The MSPE-adjusted statistic is computed as

$$\sqrt{P}[\hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - \text{adj.})]/\sqrt{\hat{V}}, \tag{4.4}$$

(see (3.1) and (3.4)), and is compared to a standard normal distribution. A second MSPE statistic is computed as

$$\sqrt{P}(\hat{\sigma}_1^2 - \hat{\sigma}_2^2)/\sqrt{\hat{W}},$$
$$\hat{W} \equiv P^{-1}\Sigma_{t=T-P+1}^{T}(d_t - \bar{d})^2, \quad d_t \equiv y_{t+1}^2 - (y_{t+1} - X_{t+1}'\hat{\beta}_t)^2,$$
$$\bar{d} \equiv P^{-1}\Sigma_{t=T-P+1}^{T}d_t. \tag{4.5}$$

and is compared to both standard normal critical values (MSPE-normal) and to McCracken's (2004) asymptotic critical values (MSPE-McCracken). (In our simulations and empirical application, $\hat{V}$ and $\hat{W}$ are virtually identical. Since $\hat{V} = \hat{W} + 2\text{cov}[\hat{f}_{t+1}, (X_{t+1}'\hat{\beta}_t)^2] - \text{var}[(X_{t+1}'\hat{\beta}_t)^2]$, this indicates that the variance of $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$ dominates that of $(X_{t+1}'\hat{\beta}_t)^2$.) The Chao, Corradi, and Swanson statistic is computed as described in their paper, with a heteroskedasticity consistent covariance matrix; it is compared to $\chi^2(2)$ critical values. In all cases, we report results based on a nominal size of 10 percent. Unreported results based on the 5 percent significance level are qualitatively similar.

Following the logic of Ashley et al. (1980) and Clark and McCracken (2001, 2003) among others, we treat all three MSPE tests as one-sided. Clark and McCracken note that because the models are nested, the alternative is that in large samples model 2 predicts better than model 1. That is, the alternative is one-sided because, if the restrictions imposed on model 1 are not true, we expect forecasts from model 2 to be superior to those from model 1.[7]

In presentation of simulation results, we will have occasion to refer to these procedures as "conventional asymptotic" ones, because they rely on asymptotic normality of the test statistic. This is in contrast to bootstrap versions of the tests. Bootstrap critical values are computing using a model-based, wild bootstrap procedure, imposing the null of no predictability of y_t. The approach is essentially that of Mark (1995) and Kilian (1999), although modified as recommended by Goncalves and Kilian (2004) to ensure accuracy in the presence of certain forms of conditional heteroskedasticity. The procedure is exactly the same for calculating bootstrap size and power.

Specifically, for each Monte Carlo data set on y_t and x_t ($R + P$ observations), we first estimate an AR(1) model for x_t, obtaining residuals $\hat{v}_t$. We then take the following steps in each of 999 bootstrap draws. (1) We generate a series of $R + P$ i.i.d. standard normal innovations η_t. (2) Imposing the null of no predictability, a

---

[7]As noted above, our MSPE-adjusted test is akin to the *t*-statistic for forecast encompassing considered by Harvey et al. (1998), West (2001) and Clark and McCracken (2001, 2003). Even in the non-nested environment considered by Harvey et al. (1998) and West (2001), under the null that the model 1 forecast encompasses model 2, the numerator of the test statistic will be less than or equal to 0. Under the alternative that model 2 contains added information, the numerator should be positive.

bootstrap series y_t^* is formed as $\eta_t y_t$. (3) A bootstrap series x_t^* is constructed using the estimated AR(1) model for x_t and innovation $\eta_t \hat{v}_t$, taking an "historical" observation on x drawn at random from the sample as the initial observation on x_0^*. (4) Using rolling forecasts of y_t^* from the null and alternative models (the latter estimated with the bootstrap data), artificial MSPE and CCS test statistics are formed. (5) Finally, using the bootstrap samples of MSPE and CCS test statistics generated by these steps, we perform our one-sided tests using as a critical value the 90th percentile of the bootstrap distribution—the 900th largest test statistic.

4.3. Simulation results: size

Size results for DGP 1 (calibrated roughly to exchange rate data) are given in Table 1, for DGP 2 (calibrated roughly to stock price data) are given in Table 2, while extensions to DGP 2 that allow conditional heteroskedasticity are in Table 3.

Table 1
Empirical size: DGP 1. Nominal size = 10%

	$P = 48$	$P = 96$	$P = 144$	$P = 240$	$P = 480$	$P = 1200$
A. $R = 60$						
MSPE-adjusted	0.073	0.072	0.072	0.074	0.079	0.085
MSPE-normal	0.007	0.002	0.000	0.000	0.000	0.000
MSPE-McCracken	0.083	0.074	0.048	0.050	0.036	0.025
CCS	0.144	0.125	0.119	0.115	0.108	0.107
B. $R = 120$						
MSPE-adjusted	0.070	0.063	0.065	0.067	0.074	0.081
MSPE-normal	0.020	0.008	0.004	0.001	0.000	0.000
MSPE-McCracken	0.089	0.087	0.079	0.073	0.063	0.052
CCS	0.144	0.121	0.119	0.114	0.108	0.100
C. $R = 240$						
MSPE-adjusted	0.079	0.072	0.070	0.067	0.066	0.072
MSPE-normal	0.034	0.020	0.014	0.004	0.001	0.000
MSPE-McCracken	0.103	0.097	0.093	0.097	0.085	0.072
CCS	0.135	0.121	0.116	0.113	0.103	0.099

Notes:
1. The predictand y_{t+1} is i.i.d. N(0, 1); the alternative model's predictor x_t follows an AR(1) with parameter 0.95; data are conditionally homoskedastic. In each simulation, one step ahead forecasts of y_{t+1} are formed from the martingale difference null and from rolling estimates of a regression of y_t on $X_t = (1, x_{t-1})'$.
2. R is the size of the rolling regression sample. P is the number of out-of-sample predictions.
3. Our MSPE-adjusted statistic, defined in (3.1) and (4.4), uses standard normal critical values. MSPE-normal, defined in (4.5), refers to the usual (unadjusted) t-test for equal MSPE, and also uses standard normal critical values. MSPE-McCracken relies on the MSPE-normal statistic but uses the asymptotic critical values of McCracken (2004).
4. The number of simulations is 10,000. The table reports the fraction of simulations in which each test rejected the null using a one-sided test at the 10% level. For example, the figure of 0.078 in panel A, $P = 48$, MSPE-adjusted, indicates that in 780 of the 10,000 simulations the MSPE-adjusted statistic was greater than 1.28.

Table 2
Empirical size: DGP 2. Nominal size = 10%

	$P = 48$	$P = 96$	$P = 144$	$P = 240$	$P = 480$	$P = 1200$
A. $R = 60$						
MSPE-adjusted	0.101	0.085	0.081	0.083	0.082	0.089
MSPE-normal	0.023	0.003	0.001	0.000	0.000	0.000
MSPE-McCracken	0.136	0.098	0.062	0.057	0.036	0.019
CCS	0.244	0.188	0.159	0.132	0.118	0.102
B. $R = 120$						
MSPE-adjusted	0.101	0.090	0.077	0.074	0.075	0.083
MSPE-normal	0.041	0.017	0.006	0.002	0.000	0.000
MSPE-McCracken	0.138	0.128	0.101	0.082	0.062	0.042
CCS	0.251	0.186	0.160	0.134	0.113	0.108
C. $R = 240$						
MSPE-adjusted	0.101	0.090	0.085	0.070	0.073	0.079
MSPE-normal	0.059	0.036	0.023	0.009	0.002	0.000
MSPE-McCracken	0.137	0.126	0.119	0.105	0.092	0.073
CCS	0.239	0.176	0.160	0.129	0.115	0.109

Notes:
1. See the notes in Table 1.
2. DGP 2 differs from DGP 1 in the variance–covariance matrix of $(y, x)'$.

Table 3
Empirical size: DGP 2 with conditional heteroskedasticity. Nominal size = 10%

	$P = 48$	$P = 96$	$P = 144$	$P = 240$	$P = 480$	$P = 1200$
A. *GARCH*						
MSPE-adjusted	0.101	0.089	0.076	0.074	0.078	0.085
MSPE-normal	0.039	0.018	0.007	0.001	0.000	0.000
MSPE-McCracken	0.137	0.127	0.099	0.081	0.067	0.055
CCS	0.255	0.191	0.158	0.132	0.118	0.109
B. *Multiplicative conditional heteroskedasticity*						
MSPE-adjusted	0.107	0.087	0.076	0.067	0.052	0.048
MSPE-normal	0.033	0.015	0.005	0.002	0.000	0.000
MSPE-McCracken	0.131	0.121	0.103	0.099	0.110	0.199
CCS	0.246	0.200	0.177	0.152	0.129	0.116

Notes:
1. See the notes in Table 1. The regression sample size R is 120.
2. In the upper panel of results, the predictand y_{t+1} is a GARCH process, with the parameterization given in Eq. (4.2). In the lower panel, the predictand y_{t+1} has conditional heteroskedasticity of the form given in Eq. (4.3), in which the conditional variance at t is a function of x_{t-1}^2.

We begin with Table 1. The ''MSPE-adjusted'' lines in the table give the performance of our proposed statistic. Since the nominal size of the test is 10%, the ideal value is 0.100. We see in Table 1 that our statistic is slightly undersized, with
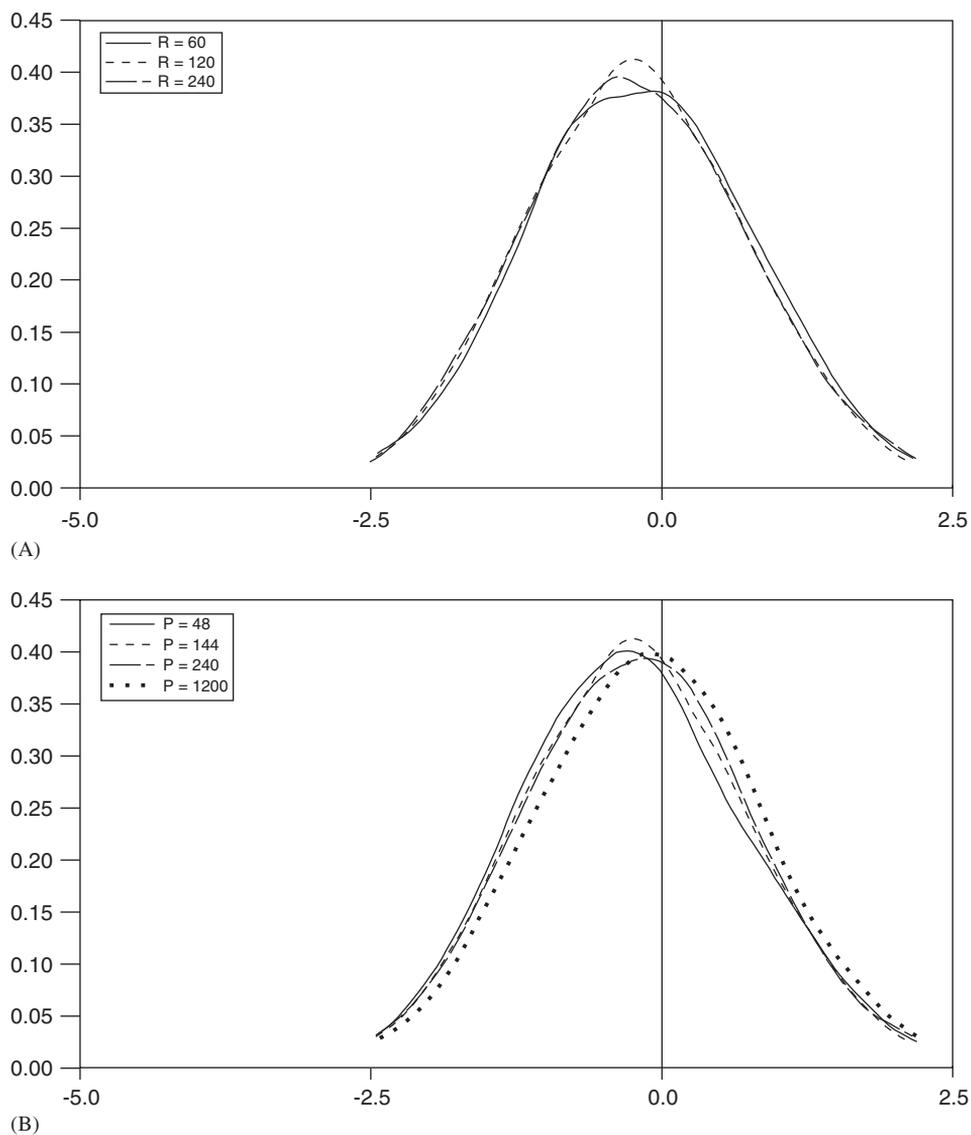
Fig. 3. Density of simulation MSPE-adjusted test statistic under the null, DGP 1: (A) R Varying, $P = 144$; (B) $R = 120$, P Varying.

actual sizes ranging from 6.3% (panel B, $P = 96$) to 8.5% (panel A, $P = 1200$). The actual size does not seem to vary systematically with R. The undersizing does seem to be reduced for larger values of P, though the pattern is mixed. Improvement with P but not with R is consistent with our asymptotic theory, which holds R fixed but assumes $P \to \infty$.

These features are illustrated in Fig. 3, which presents smoothed density plots of distributions of MSPE-adjusted statistics. The ideal figure is a N(0, 1) distribution. The top panel presents results for all three values of R, with P held at 144. We see that the undersizing documented in Table 1 results because the distributions are shifted slightly to the left. (Recall that we are examining one-sided tests in which one rejects only when the statistic is above +1.28.) The leftward shift is a small sample phenomenon not captured by our asymptotic theory. There is no apparent pattern for empirical sizes to either improve or degrade with R. The bottom panel holds R fixed at 120 and lets P vary over the values 48, 144, 240 and 1200. Again the figures are shifted slightly to the left. But there is a tendency for the downward shift to lessen as P increases: the distribution for $P = 1200$ is distinctly better centered than are those for smaller values of P.

The performance of the MSPE statistic that uses standard normal values ("MSPE-normal" in the tables) is not as good. This statistic results in far too few rejections. Indeed, the empirical size of the test is essentially 0 in most instances, especially for larger values of P. For example, we see in panel B of Table 1 that when $R = 120$, comparing the MSPE statistic against the standard normal distribution produces a nominal size of 2.0 percent for $P = 48$ and 0.8 percent or less for all values of P greater than 48. For given P, the performance of the test improves monotonically as R gets larger, for example increasing for $P = 144$ from 0.000 ($R = 60$) to 0.004 ($R = 120$) to 0.014 ($R = 240$). As well, performance degrades as P gets bigger, for example falling for $R = 120$ from 0.008 ($P = 96$) to 0.001 ($P = 240$) to 0.000 ($P = 1200$).

To help to interpret the results for MSPE-normal, Fig. 4 presents smoothed density plots of MSPE-normal corresponding to some of the simulations in Table 1. Variation with R for fixed $P = 144$ is depicted in the panel A, while variation with P for fixed $R = 120$ is depicted in the panel B. We see that for all values of P and R, the test is undersized because the distribution is markedly shifted leftwards. The leftward shift reflects the parameter noise in the alternative model forecast, and is predicted by our asymptotic theory. See Eq. (2.8) and panel A in Figs. 1 and 2. Improvement in size as R increases is clear in panel A. As noted in Section 2, this follows because parameter noise shrinks as the number of regression observations R rises. The degradation in performance as P increases is equally clear in panel B. This, too, follows from our asymptotic approximation. The leftward shifts as P increases occur because the scaling of the difference in MSPEs by $\sqrt{P}$—scaling associated with making the asymptotic distribution non-degenerate—results in the $\sqrt{P}$-scaled parameter noise component of the difference in MSPEs rising with P. The parameter noise term $P^{-1}\Sigma_{t=T-P+1}^{T}(X'_{t+1}\hat{\beta}_t)^2$ converges to a constant, but $\sqrt{P}$ times the same term grows with P.

Now turn to a third statistic, "MSPE-McCracken". Since this statistic is identical to MSPE-normal, its density is also represented by Fig. 4. But this statistic uses McCracken's (2004) critical values, which asymptotically ($P \to \infty$ and $R \to \infty$) account for the leftward shifts depicted in Fig. 4. We see in Table 1 that comparing the unadjusted t-statistic for equal MSPE against those critical values can also yield a well-sized test in many, although not all, circumstances. For smaller values of P,
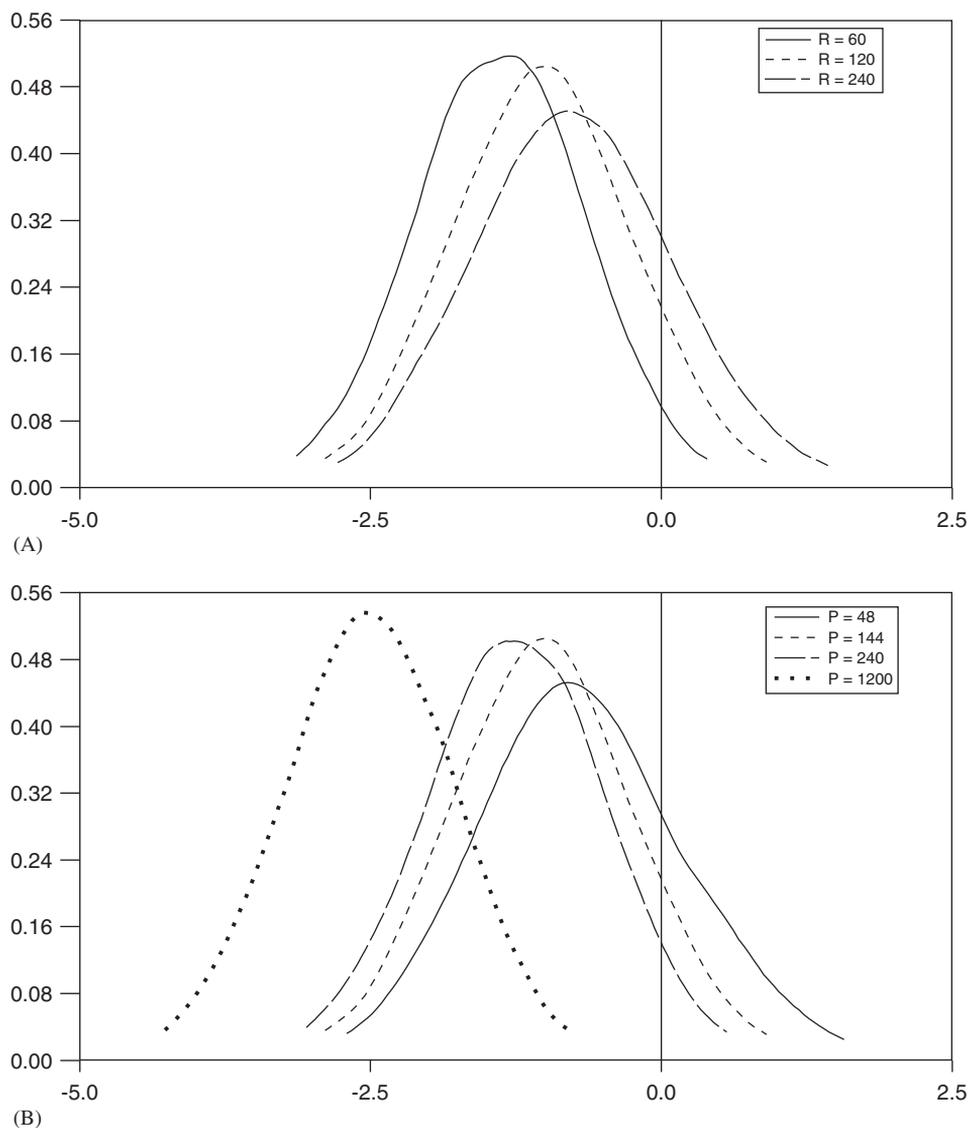
Fig. 4. Density of simulation MSPE test statistic (unadjusted) under the null, DGP 1: (A) R Varying, $P = 144$; (B) $R = 120$, P Varying.

using McCracken's critical values yields a test that is usually but not always slightly undersized. With $R = 120$ and $P = 96$, for example, the empirical size is 8.7 percent; for $R = 120$ and $P = 1200$, the size is 5.2 percent. Some unreported simulations show these size distortions are ameliorated when R and P are *both* bigger. Apparently, in our rolling forecast approach, for large values of P and P/R, both P and R may need

to be large for the finite sample distribution to closely match McCracken's asymptotic approximation.

The final statistic tabulated in Table 1 is that of Chao et al. (2001) ("CCS" in the tables). This statistic is well-sized in larger forecast samples (larger P), but tends to be oversized in smaller samples. For example, for $R = 120$, the sizes are 10.0% for $P = 1200$ but 14.4% for $P = 48$.

Simulation results for DGP 2 are presented in Tables 2 (homoskedastic data) and 3 (conditionally heteroskedastic data). These tell a broadly similar story. Briefly, performance of MSPE-adjusted continues to be good; performance of MSPE-normal continues to be poor but not perceptibly poorer than in Table 1; performance of both MSPE-McCracken and CCS declines. The oversizing of CCS is congruent with the finding in a number of in-sample studies that regression tests of return predictability tend to be oversized for data similar to ours (e.g., Nelson and Kim, 1993).

To take the statistics in reverse order of the rows of the panels of Tables 2 and 3: the tendency of CCS to overreject for smaller P is now quite marked, with sizes over 20% found for $P = 48$ for all panels in Tables 2 and 3, though good performance for $P = 1200$ continues to be displayed. The size distortions for smaller P appear to be due to finite-sample imprecision in the (heteroskedasticity-robust) variance estimate that enters the CCS statistic. In results not reported in the tables, we found that imposing the null hypothesis or imposing homoskedasticity in simulated data that are homoskedastic improves the variance estimate and greatly reduces the size problems; in all cases, adding more regressors to the forecasting model exacerbates the size problems. Ultimately, in small samples, there appears to be some (size) advantage to our approach of projecting the null forecast error onto the scalar $X'_{t+1}\hat{\beta}_t$ rather than the CCS approach of projecting the null error onto the vector X_{t+1}. That advantage appears to be due to the greater precision of estimates of the variance of the scalar compared to estimates of the variance of the vector.

"MSPE-McCracken" is more variable in Tables 2 and 3 than in Table 1. It shows a tendency to be modestly oversized for small P and modestly undersized for large P. Performance is weaker for conditionally heteroskedastic data (Table 3), no doubt because McCracken's critical values assume conditional homoskedasticity.[8]

"MSPE-normal" continues to be egregiously undersized, particularly for large P (size of 0.0% for $P = 1200$ in every single panel of Tables 1–3) and small R (all sizes at or below 2.3% for $R = 60$).

Finally, performance of MSPE-adjusted in Tables 2 and 3 is similar to that in Table 1. The test generally is modestly undersized. The undersizing does, however, tend to be less modest in Tables 2 and 3, with sizes ranging from around 7% to around 10% (as opposed to Table 1's 6–8%). Once again, performance does not vary systematically with R. The figures for DGP 2 do not display a clear tendency for

---

[8]Clark and McCracken (2003) develop an asymptotic approximation in which heteroskedasticity makes the limiting null distribution of the (unadjusted) test for equal MSPE dependent on the parameters of the data generating process. The results in Clark and McCracken (2003) indicate that, with heteroskedasticity, a simple bootstrap approach to inference will yield accurately sized tests.

performance to improve with larger P, especially for data with multiplicative conditional heteroskedasticity. In results not reported in the table, we experimented with larger sample sizes for such data. Specifically, for $R = 120$, we found a size of 7.3% for $P = 12,000$ and 8.2% for $P = 24,000$. This is an improvement over the size of 4.9% for $P = 1200$ reported in panel B of Table 3, but also indicates that for such heteroskedasticity very large sample sizes are required for the asymptotic approximation to work well.

Our basic one step ahead results carry over to long horizon forecasts. Here are some representative results for DGP 1, for $R = 120$, $P = 144$:

	Horizon τ				
	1	6	12	24	36
MSPE − adjusted	0.065	0.065	0.071	0.076	0.078
MSPE − normal	0.004	0.004	0.006	0.010	0.015
CCS	0.119	0.098	0.096	0.098	0.107

$$(4.6)$$

(We do not report results for MSPE-McCracken because McCracken has supplied critical values only for one step ahead forecasts. See Clark and McCracken (2003) for a bootstrap approach to critical values for long horizon forecasts.)

Our basic results roughly continue to hold in the presence of modest misspecification. By "modest" we mean that departures from a martingale difference are sufficiently small that even a careful investigator might wrongly assume a martingale difference. We investigate this possibility by positing that y_t follows an MA(1) process with a small first order autocorrelation:

$$y_t = e_t + u_t + 0.5u_{t-1}. \tag{4.7}$$

In (4.7), e_t and u_t are i.i.d. normal variables, with $\sigma_e^2 = 0.8$ and σ_u^2 chosen so that $\sigma_y^2 = 1$. The implied first order autocorrelation of y_t is 0.08 (and of course all other autocorrelations are zero). We consider this a modest departure from the random walk model. Other parameters are as in DGP 1. We repeat the analysis above for $R = 120$. In particular, we assume that the investigator uses a sample variance (e.g., the formula for $\hat{V}$ in (3.4)) when one should account for serial correlation in y_t by using an estimate of the long run variance instead. Results are as follows:

	$P = 48$	$P = 96$	$P = 144$	$P = 240$	$P = 480$	$P = 1200$
MSPE − adjusted	0.095	0.087	0.089	0.100	0.120	0.161
MSPE − normal	0.027	0.011	0.005	0.001	0.000	0.000
MSPE − McCracken	0.108	0.107	0.100	0.098	0.094	0.091
CCS	0.181	0.154	0.151	0.142	0.146	0.144

$$(4.8)$$

Rejection rates increase for all tests, essentially because with the DGP (4.7), the relevant variance is smaller than the long run variance. The increase is modest, because the divergence between variance and long run variance is small.

Naturally, one may wish to guard against such possibilities by using an estimator of the long run variance. We have not investigated in detail the effect of using such an estimator. We did, however, investigate alternative estimators of the long run variance in connection with the long horizon results reported in (4.6). We found that inference is extraordinarily sensitive to the estimator used to account for the serial correlation in the forecast. As noted above, we used the estimator of West (1997) and Hodrick (1992). Use of Newey and West (1994) or Andrews (1991) resulted in substantially poorer performance. For example, for CCS, use of a quadratic spectral kernel with bandwidth chosen as recommended in Andrews (1991) resulted in the following sizes for horizons of 6, 12, 24 and 36 periods: 0.324, 0.427, 0.579, and 0.679.

Overall, the results discussed in this section show that in applications in which the null model is a martingale difference, it is possible, with a simple adjustment, to use a standard distribution in conducting a reliable test of equal MSPE.

4.4. Simulation results: bootstrapping

Table 4 presents results on bootstrap size. For concision, we report results only for $R = 120$, DGPs 1 and 2. Full results are available in the not for publication appendix. Since MSPE-normal and MSPE-McCracken use the same test statistic

Table 4
Bootstrap size. $R = 120$, Nominal size = 10%

	$P = 48$	$P = 96$	$P = 144$	$P = 240$	$P = 480$	$P = 1200$
A. Size, DGP 1						
MSPE-adjusted, std. normal	0.070	0.063	0.065	0.067	0.074	0.081
MSPE-adjusted, bootstrap	0.085	0.088	0.095	0.094	0.093	0.096
MSPE, bootstrap	0.090	0.094	0.101	0.101	0.099	0.100
CCS, bootstrap	0.111	0.105	0.107	0.106	0.105	0.099
B. Size, DGP 2						
MSPE-adjusted, std. normal	0.101	0.090	0.077	0.074	0.075	0.083
MSPE-adjusted, bootstrap	0.104	0.103	0.096	0.096	0.094	0.097
MSPE, bootstrap	0.106	0.107	0.100	0.098	0.097	0.099
CCS, bootstrap	0.153	0.128	0.120	0.109	0.098	0.099

Notes:
1. The entries for "MSPE-adjusted, std. normal" repeat the "MSPE-adjusted" results reported in Table 1 (panel A) and Table 2 (panel B).
2. The "bootstrap" entries rely on the percentile-*t* method, using the wild bootstrap of Goncalves and Kilian (2004). The number of simulations is 10,000. The number of bootstrap replications per simulation is 999. See text for additional details.
3. See the notes in Table 1.

(4.5), bootstrap results are identical for the two and the tables report a single value in lines labeled "MSPE, bootstrap".

The first row of panels A and B of Table 4 ("MSPE-adjusted, normal") repeats the corresponding entries from Tables 1 and 2, for convenience. The next three lines present sizes when hypothesis tests are conducted by the bootstrap described above. These three lines indicate that the bootstrap generally does a good job. It lessens the mild size distortions of our procedure and the very marked size distortions of MSPE-normal. And the last lines in these panels ("CCS, bootstrap") indicate that apart from when P is very small (e.g., $P = 48$, panel B), bootstrapping of CCS also eliminates size distortions.

Thus, bootstrapping produces modest improvements in size relative to conventional asymptotics. (Indeed, our bootstrap results might be better than can be expected in practice, since for simplicity our bootstrap procedure generously but unrealistically allowed the investigator to know that x_t was generated by an AR(1) [though of course with unknown parameter].) We defer summary and further discussion of bootstrap results to first present results on power.

Table 5
Size adjusted and bootstrap power. $R = 120$, Size $= 10\%$

	$P = 48$	$P = 96$	$P = 144$	$P = 240$	$P = 480$	$P = 1200$
A. Size adjusted, DGP 1						
MSPE-adjusted	0.345	0.464	0.554	0.677	0.854	0.984
MSPE	0.283	0.408	0.503	0.643	0.842	0.982
CCS	0.232	0.374	0.483	0.672	0.913	0.999
B. Bootstrap, DGP 1						
MSPE-adjusted	0.315	0.445	0.545	0.666	0.846	0.984
MSPE	0.268	0.394	0.500	0.645	0.842	0.981
CCS	0.249	0.381	0.495	0.679	0.917	0.999
C. Size Adjusted, DGP 2						
MSPE-adjusted	0.114	0.120	0.127	0.138	0.147	0.171
MSPE	0.119	0.122	0.131	0.142	0.158	0.184
CCS	0.104	0.114	0.125	0.142	0.203	0.351
D. Bootstrap, DGP 2						
MSPE-adjusted	0.118	0.123	0.122	0.133	0.141	0.165
MSPE	0.127	0.133	0.132	0.139	0.152	0.176
CCS	0.163	0.148	0.146	0.156	0.199	0.350

Notes:
1. The DGP is defined in Eq. (4.1), with: $b = -2$ (panels A and B) or $b = 0.365$ (panels C and D); $e_{t+1} \sim$ i.i.d. N(0, 1); $x_t \sim$ AR(1) with parameter 0.95; data are conditionally homoskedastic. In each simulation, one step ahead forecasts of y_{t+1} are formed from the martingale difference null and from rolling estimates of a regression of y_t on $X_t = (1, x_{t-1})'$.
2. In panels A and C, power is calculated by comparing the test statistics against simulation critical values, calculated as the 90th percentile of the distributions of the statistics in the corresponding size experiment reported in Table 1 (panel A) or Table 2 (panel C).
3. For description of the bootstrap procedure used in panels B and D, see the text and notes in Table 4.

4.5. Simulation results: power

Table 5 presents results on power. We studied the finite-sample power of the MSPE-adjusted test, the simple (unadjusted) MSPE test, and the CCS test, using both empirical critical values from our initial size experiments and bootstrap critical values, and for a range of values of R. For conciseness, we report only the results for $R = 120$ and homoskedastic data. We calculate size-adjusted power in experiments involving DGP 1 using critical values taken from the corresponding size experiment for DGP 1. The DGP is as given in (4.1) with $b = -2$, a value consistent with the exchange rate data considered in the next section. Similarly, we calculate size-adjusted power for DGP 2 experiments using critical values from the size experiments with DGP 2. Here, the DGP is as given in (4.1) with $b = 0.365$. For these size adjusted tests, the actual size is 10 percent. For each DGP, we calculate bootstrap power using 10 percent critical values with the procedure described in Section 4.2.

Overall, the powers of the tests are similar. The most notable exception is that in DGP 2, CCS has distinctly more power than MSPE or MSPE-adjusted when $P = 1200$. As well, in DGP 1, CCS perhaps has slightly less power than MSPE-adjusted for small P, slightly more power for large P. In results not reported in Table 5, we found that given large P, the power difference between the CCS and MSPE-adjusted tests falls as R rises. This pattern (which is mirrored in the absence of a relationship between the power of the CCS test and R (given P)), reflects the fact that the MSPE-adjusted test is affected by parameter estimation noise while the CCS test is not. The CCS test is not affected because the test is based on simply projecting the martingale difference y_{t+1} on X_{t+1} (while, as noted above, MSPE-adjusted projects y_{t+1} onto $X'_{t+1}\hat{\beta}_t$).[9]

In results reported in the appendix, we find, as did Clark and McCracken (2001, 2003), that size-adjusted power of both the adjusted and unadjusted MSPE tests tends to rise with R (given P) and with P (given R). The latter is to be expected; the positive relationship between power and R reflects the increased precision of the forecasting model parameter estimates.

Results presented so far relate to size-adjusted power. But to interpret empirical results presented in the next section, it may be useful to explicitly present some numbers on power that are not size adjusted. In particular, let us consider the probability of rejecting the null under DGP 1 with $b = -2$, a configuration chosen because of its relevance for our empirical work. Here are *raw* (as opposed to

---

[9]In unreported results, we also examined the performance of the *F*-type test of equal MSPE developed by McCracken (2004) and the *F*-type test of encompassing proposed by Clark and McCracken (2001). In general, with homoskedastic data, these additional tests have good size properties. The behavior of the *F*-test for equal MSPE is comparable to the reported behavior of the *t*-test for equal MSPE compared to McCracken's (2004) critical values. The encompassing test is very close to correctly sized in DGP 1 experiments and somewhere between correctly sized (larger P) and slightly oversized (smaller P) in DGP 2 experiments. In general, the size-adjusted powers of these additional tests are broadly comparable to that of the MSPE-adjusted test. For larger values of R, though, there tend to be more noticeable differences among the powers of the tests, in line with Clark and McCracken (2001, 2003).

size-adjusted) probabilities of rejecting at the 10 percent level in simulations with sample sizes most relevant to our empirical example:

	$R = 120, P = 144$	$R = 120, P = 240$
MPSE − adjusted	0.480	0.614
MSPE − normal	0.086	0.086
MSPE − McCracken	0.457	0.588
CCS	0.518	0.694

$$(4.9)$$

The values in Eq. (4.9) illustrate two points. The first is that the downward shift in MSPE-normal makes this statistic ill-suited to detect departures from the null. The second is that in our simulated data, which we believe to be plausibly calibrated, the probability of detecting the failure of the null is substantial (around 0.5 or 0.6) but by no means overwhelming—a sobering, but important, reminder that asset price data may be sufficiently noisy that it will be difficult to detect departures from martingale difference behavior with sample sizes currently available.

The results on unadjusted power also are helpful in interpreting the empirical results presented in the next section. In these results, we sometime find that the martingale difference model has a smaller MSPE than the alternative, but our adjusted MSPE statistic nonetheless rejects the null at conventional significance levels. In our simulations under the alternative, we do indeed find substantial positive probability of this occurring. For example, in the $R = 120$, $P = 240$ simulation reported in (4.9), $\hat{\sigma}_1^2 < \hat{\sigma}_2^2$ in about 47% of the simulations, but as indicated in (4.9) we reject at the 10 percent level in about 61% of the simulations. We examined this seemingly peculiar finding (martingale difference has lower MSPE, but still can be rejected at conventional significance levels) over a range of values of R and P. Unsurprisingly, we found that increases in either R or P led to decreases in the fraction of the time in which the martingale difference model had a smaller MSPE and to increases in power. For example, with $R = 120$, $P = 1200$ the martingale difference model had a smaller MSPE in about 33% of the simulations and the power of our test was about 98%. Thus one can interpret the finding as indicating that there is indeed a correlation between y_t and the predictors used by the alternative model, but sample sizes are too small for the alternative model to reliably outpredict the martingale difference model.[10]

Let us close with a comparison of the conventional asymptotic and bootstrap implementations of our proposed statistic. Overall, the bootstrap implementation shows modestly better size. Table 5 indicates that bootstrap power is a slightly worse than size-adjusted power; a comparison of Table 5 and Eq. (4.9) indicates that bootstrap power is slightly better than unadjusted power. Bootstrapping involves a

---

[10]Or more precisely, sample sizes are too small for the alternative to reliably outpredict using a least squares estimator. We have not experimented with predictions that rely on other schemes for exploiting the correlation between y_t and X_t.

trivial cost in terms of computation (at least when the alternative model is linear) and what we view as a more substantial cost in terms of programming time. (No, we do not agree with what we sense is the view of many economists that as long as computation time is cheap there is no reason to be hesitant to use simulation based procedures. It has been our experience that producing carefully debugged code is time consuming.) We invite readers to trade off these additional costs of bootstrapping against the modest improvement in size and modest, mixed effects on power. Our own utility functions are such that the conventional asymptotic version of our statistic will often be preferred.

5. Empirical example

To illustrate our approach, we apply the MSPE-adjusted, MSPE-normal, MSPE-McCracken and CCS tests to monthly forecasts of four US dollar bilateral exchange rates, for Canada, Japan, Switzerland, and the U.K. Under the null, the exchange rate follows a random walk. Our alternative model, based on interest parity, relates the change in the spot exchange rate from the end of month t to the end of month $t+1$ to a constant and the one-month interest rate differential at the end of month t. (The end-of-month values are defined as those on the last business day of the month.) The exchange rate data were obtained from the Board of Governor's FAME database; the one-month interest rates, which are averages of bid and ask rates on Eurocurrency deposits (London close), were obtained from Global Insight's FACS database. We use a rolling window of 120 observations to estimate the alternative forecast model. Reflecting data availability, for Canada and Japan, we examine forecasts from January 1990 to October 2003 ($P = 166$, with the first regression sample running from January 1980 to December 1989); for Switzerland and the U.K., the forecast sample is January 1985–October 2003 ($P = 226$, with the first regression sample running from January 1975 to December 1984).

While unreported in the interest of brevity, estimates of our interest parity regressions share the basic features highlighted in the long literature on uncovered interest parity (see, for example, the surveys of Taylor (1995) and Engel (1996) and studies such as Hai et al. (1997) and Backus et al. (2001)): significantly negative slope coefficients that wander substantially over the sequence of rolling samples (from -3.9 to -0.2, with a median of about -1.8, taking the yen as representative) and low regression R^2s (from essentially zero to 5.2%, with a median of about 2%, again for the yen). Simple uncovered parity of course implies the coefficient should be 1. But even if the data violate uncovered parity, interest rate differentials could have predictive content for exchange rates. Accordingly, we follow a number of studies (see Clarida and Taylor, 1997, for a recent example) in evaluating forecasts based on interest differentials.

Table 6 contains our results. The table reflects the widely known difficulty of beating, in MSPE, a random walk model of exchange rates: for only one of the four countries, Canada, does the interest parity model yield a forecast with MSPE lower than that of the random walk forecast. Even for Canada, the advantage of the parity

Table 6
Forecasts of monthly changes in US Dollar exchange rates

(1) Country	(2) Prediction sample	(3) $\hat{\sigma}_1^2$	(4) $\hat{\sigma}_2^2$	(5) adj.	(6) $\hat{\sigma}_2^2$-adj.	(7) MSPE-adjusted $\hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - \text{adj.})$	(8) MSPE-normal $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$	(9) CCS
Canada	1990:1–2003:10	2.36	2.32	0.09	2.22	0.13 (0.08)	0.04	
						1.78**	0.54††	3.67
Japan	1990:1–2003:10	11.32	11.55	0.75	10.80	0.53 (0.43)	−0.23	
						1.24	−0.52	5.23*
Switzerland	1985:1–2003:10	12.27	12.33	0.96	11.37	0.90 (0.48)	−0.06	
						1.88**	−0.13	2.43
U.K.	1985:1–2003:10	9.73	10.16	0.44	9.72	0.01 (0.33)	−0.43	
						0.03	−1.27	0.78

Notes:

1. In column (3), $\hat{\sigma}_1^2$ is the out of sample MSPE of the "no change" or random walk model, which forecasts a value of zero for the one month ahead change in the exchange rate.

2. In column (4), $\hat{\sigma}_2^2$ is the out of sample MSPE of a model that regresses the exchange rate on a constant and the previous month's cross-country interest differential. The estimated regression vector and the current month's interest differential are then used to predict next month's exchange rate. Rolling regressions are used, with a sample size R of 120 months.

3. In column (5), "adj." is the adjustment term $P^{-1}\Sigma_{t=T-P+1}^{T}(X'_{t+1}\hat{\beta}_t)^2$, where: P is the number of predictions, $P = 166$ for Canada and Japan, $P = 226$ for Switzerland and the U.K.; $T = 2003{:}10$; $X_{t+1} = $ (constant, interest differential at end of month t)$'$; $\hat{\beta}_t$ is the estimated regression vector.

4. In columns (7)–(9), standard errors are in parentheses and t-statistics (columns (7) and (8)) or a $\chi^2(2)$ statistic (column (9)) are in italics. Standard errors are computed as described in the text. * and ** denote test statistics significant at the 10 and 5 percent level, respectively, based on one sided tests using critical values from a standard normal (columns (7) and (8)) or chi-squared (CCS) distribution. In columns (7) and (8), † and †† denote statistics significant at the 10 and 5 percent level based on McCracken's (2004) asymptotic critical values.

5. Data are described in the text. See notes to earlier tables for additional definitions.

model over the random walk is slight. Accordingly, without any adjustment, comparing the simple MSPE test against standard normal critical values systematically—that is, for all countries—fails to reject the null random walk. For Canada, though, comparing the simple MSPE test against McCracken's (2004) critical values does lead to a rejection of the null random walk.

The results of our adjustment procedure highlight the potential for noise associated with estimation of the alternative model's parameters to create an upward shift in the model's MSPE large enough that the random walk has a lower MSPE even when the alternative model is true. The estimated adjustments in column (5) in Table 6 correspond to the term $P^{-1}\Sigma_{t=T-P+1}^{T}(X'_{t+1}\hat{\beta}_t)^2$. These range from 0.09 for Canada to 0.96 for Switzerland, corresponding to about 4 percent of the alternative model's MSPE for Canada and 7.8 percent for Switzerland. For Canada,

the adjustment makes the advantage of the interest parity model over the random walk look more substantial than it does on the basis of the unadjusted difference in MSPEs. For the other three countries, the adjustment is large enough to make the adjusted difference in MSPEs positive even though the unadjusted difference is negative. And for Switzerland as well as Canada, the MSPE-adjusted test rejects the null random walk; for Japan, the CCS test also rejects the null. The MSPE-normal test does not reject for Canada, nor, of course, for the three countries for which the MSPE of the random walk model was less than the unadjusted MSPE of the interest differential model.

Thus, while the unadjusted MSPE test would seem to provide uniform support for the random walk null, our MSPE-adjusted test, which adjusts for the parameter noise in the alternative model, provides some evidence in favor of an alternative model that includes the interest rate differential. That is, even though parameter estimation noise may cause a random walk forecast to have a lower MSPE, there is some evidence that exchange rate changes are correlated with lagged interest differentials. Such results highlight both the practical importance of taking parameter noise into account in conducting tests of equal MSPE in nested models and the potential value of our simple procedure for doing so.

6. Conclusion

If the martingale difference hypothesis holds, out-of-sample mean squared prediction errors should be smaller for a ''no change'' model than for an alternative model that forecasts with an estimated regression vector. This is because the forecasts of the alternative model are expected to be shifted upward by noisy estimates of a regression vector whose population value is zero. We show how to adjust for the upward shift. Simulations indicate that adjustment for the shift results in hypothesis tests that are well sized. Failure to adjust results in hypothesis tests that are very poorly sized.

We have studied a problem that is relatively simple but quite important in practice. For expositional convenience, we have assumed stationarity and a parametric linear alternative. Both restrictions are easily relaxed. Substantive priorities for future work include extensions to nested model comparisons in which forecasts for the null model rely on an estimated regression vector and sampling schemes in which forecasts are generated recursively rather than with rolling regressions.

References

Andrews, D.W.K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817–858.

Ashley, R., Granger, C.W.J., Schmalensee, R., 1980. Advertising and aggregate consumption: an analysis of causality. Econometrica 48, 1149–1168.

Backus, D.K., Foresi, S., Telmer, C.I., 2001. Affine term structure models and the forward premium anomaly. Journal of Finance 56, 279–304.

Campbell, J.Y., 2001. Why long horizons? a study of power against persistent alternatives. Journal of Empirical Finance 8, 459–491.

Chao, J., Corradi, V., Swanson, N.R., 2001. Out-of-sample tests for granger causality. Macroeconomic Dynamics 5, 598–620.

Cheung, Y.-W., Chinn, M.D., Pascual, A.G., 2003. Empirical exchange rate models of the nineties: are any fit to survive? Manuscript, University of California at Santa Cruz.

Christiano, L.J., 1989. P^*: Not the inflation forecaster's holy grail. Federal Reserve Bank of Minneapolis Quarterly Review 13, 3–18.

Clarida, R.H., Sarno, L., Taylor, M.P., Valente, G., 2003. The out-of-sample success of term structure models as exchange rate predictors: a step beyond. Journal of International Economics 60, 61–83.

Clarida, R.H., Taylor, M.P., 1997. The term structure of forward exchange premiums and the forecastability of spot exchange rates: correcting the errors. Review of Economics and Statistics 79, 353–361.

Clark, T.E., McCracken, M.W., 2001. Tests of equal forecast accuracy and encompassing for nested models. Journal of Econometrics 105, 85–110.

Clark, T.E., McCracken, M.W., 2003. Evaluating long horizon forecasts. Manuscript, University of Missouri.

Clark, T.E., McCracken, M.W., 2004. Improving forecast accuracy by combining recursive and rolling forecasts. Manuscript, University of Missouri.

Corradi, V., Swanson, N.R., 2002. A consistent test for nonlinear predictive ability. Journal of Econometrics 110, 353–381.

de Jong, R.M., 1996. The Bierens test under data dependence. Journal of Econometrics 72, 1–32.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. Journal of Business and Economic Statistics 13, 253–263.

Engel, C., 1996. The forward discount anomaly and the risk premium: a survey of recent evidence. Journal of Empirical Finance 123–192.

Giacomini, R., White, H., 2003. Tests of conditional predictive ability. Manuscript, University of California at San Diego.

Goncalves, S., Kilian, L., 2004. Bootstrapping autoregressions with conditional heteroskedasticity of unknown form. Journal of Econometrics 123, 89–120.

Goyal, A., Welch, I., 2003. Predicting the equity premium with dividend ratios. Management Science 49, 639–654.

Granger, C.W.J., Newbold, P., 1977. Forecasting Economic Time Series. Academic Press, New York.

Hai, W., Mark, N.C., Wu, Y., 1997. Understanding spot and forward exchange rate regressions. Journal of Applied Econometrics 12, 715–734.

Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1054.

Harvey, D.I., Leybourne, S.J., Newbold, P., 1998. Tests for forecast encompassing. Journal of Business and Economic Statistics 16, 254–259.

Hodrick, R.J., 1992. Dividend yields and expected stock returns: alternative procedures for inference and measurement. Review of Financial Studies 5, 357–386.

Ing, C.-K., 2003. Multistep prediction in autoregressive processes. Econometric Theory 19, 254–279.

Inoue, A., Kilian, L., 2004. In-sample or out-of-sample tests of predictability: which one should we use? Econometric Reviews 23, 371–402.

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

Kilian, L., Taylor, M.P., 2003. Why is it so difficult to beat the random walk forecast of exchange rates? Journal of International Economics 60, 85–107.

Kunitomo, N., Yamamoto, T., 1985. Properties of predictors in misspecified autoregressive time series models. Journal of the American Statistical Association 80, 941–949.

Mankiw, N.G., Shapiro, M.D., 1986. Do we reject too often? Small sample properties of tests of rational expectations models. Economics Letters 20, 139–145.

Mark, N., 1995. Exchange rates and fundamentals: evidence on long-horizon predictability. American Economic Review 85, 201–218.

McCracken, M.W., 2004. Asymptotics for out of sample tests of causality. Manuscript, University of Missouri.

McCracken, M.W., Sapp, S., 2005. Evaluating the predictability of exchange rates using long horizon regressions. Journal of Money, Credit and Banking 37, 473–494.

Meese, R.A., Rogoff, K., 1983. Empirical exchange rate models of the seventies: do they fit out of sample? Journal of International Economics 14, 3–24.

Meese, R.A., Rogoff, K., 1988. Was it real? The exchange rate—interest differential over the modern floating rate period. Journal of Finance 43, 933–948.

Nelson, C.R., Kim, M.J., 1993. Predictable stock returns: the role of small sample bias. Journal of Finance 48, 641–661.

Newey, W.K., West, K.D., 1994. Automatic lag selection in covariance matrix estimation. Review of Economic Studies 61, 631–654.

Stambaugh, R.F., 1999. Predictive regressions. Journal of Financial Economics 54, 375–421.

Tauchen, G., 2001. The bias of tests for a risk premium in forward exchange rates. Journal of Empirical Finance 8, 695–704.

Taylor, M.P., 1995. The economics of exchange rates. Journal of Economic Literature 33, 13–47.

Theil, H., 1971. Principles of Econometrics. Wiley, New York.

West, K.D., 1996. Asymptotic inference about predictive ability. Econometrica 64, 1067–1084.

West, K.D., 1997. Another heteroskedasticity and autocorrelation consistent covariance matrix estimator. Journal of Econometrics 76, 171–191.

West, K.D., 2001. Tests of forecast encompassing when forecasts depend on estimated regression parameters. Journal of Business and Economic Statistics 19, 29–33.

West, K.D., Edison, H.J., Cho, D., 1993. A utility based comparison of some models of exchange rate volatility. Journal of International Economics 35, 23–46.

West, K.D., McCracken, M.W., 1998. Regression based tests of predictive ability. International Economic Review 39, 817–840.