JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **14:** 491–510 (1999)

EXCHANGE RATES AND MONETARY FUNDAMENTALS: WHAT DO WE LEARN FROM LONG-HORIZON REGRESSIONS?

LUTZ KILIAN*
Department of Economics, University of Michigan, Ann Arbor, MI 48109-1220, USA

SUMMARY

The use of a new bootstrap method for small-sample inference in long-horizon regressions is illustrated by analysing the long-horizon predictability of four major exchange rates, and the findings are reconciled with those of an earlier study by Mark (1995). While there is some evidence of exchange rate predictability, contrary to earlier studies, no evidence is found of higher predictability at longer horizons. Additional evidence is presented that the linear VEC model framework underlying the empirical study is likely to be misspecified, and that the methodology for constructing bootstrap *p*-values for long-horizon regression tests may be fundamentally flawed. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

Long-horizon regression tests are widely used in empirical finance as tests of market efficiency. They have been used, for example, in exchange rate prediction (e.g. Mark, 1995; Chinn and Meese, 1995), in the analysis of dividend yields and expected stock returns (e.g. Fama and French, 1988; Campbell and Shiller, 1988) and in studies of the term structure of interest rates (e.g. Fama and Bliss, 1987; Cutler, Poterba, and Summers, 1991).[1] In the absence of market efficiency, deviations of asset prices from their long-run equilibrium value should help predict cumulative future asset returns. Regression tests of this hypothesis typically find strong evidence of predictability at long forecast horizons, but cannot reject the null of unpredictable asset returns at short forecast horizons. This finding is often interpreted as evidence of increasing power at higher forecast horizons. However, there exists a large body of literature which questions this interpretation of long-horizon regression test results. For example, Mankiw, Romer, and Shapiro (1991), Hodrick (1992), Nelson and Kim (1993), Bollerslev and Hodrick (1995), Berkowitz and Giorgianni (1997), Bekaert, Hodrick, and Marshall (1997), and Kirby (1997) have documented that conventional long-horizon regression tests are biased in favour of finding predictability. Severe size distortions may arise from spurious regression fits and from small-sample bias in the estimates of regression coefficients and asymptotic standard errors. Previous attempts to mitigate these size distortions have only been partially successful. In this paper, a new bootstrap method for small-sample inference in long-horizon regressions is introduced. Monte Carlo evidence shows that this bootstrap test is indeed reasonably accurate in realistic situations.

* Correspondence to: L. Kilian, Dept of Economics, University of Michigan, Ann Arbor, MI 48109-1220, USA. E-mail: lkilian@umich.edu

[1] Campbell and Shiller (1988) and Mankiw, Romer, and Shapiro (1991) discuss the close relationship between long-horizon regression tests and volatility tests.

CCC 0883-7252/99/050491–20$17·50

Received 19 January 1998
Revised 1 November 1998

The use of this bootstrap method is illustrated by analysing the question of exchange rate predictability under the current float. While the standard monetary exchange rate model is clearly rejected by the data for any given period, many economists consider it a reasonable description of the long run. The monetary model predicts that at least in the long run the exchange rate must revert to its equilibrium value. As a result, current deviations from the equilibrium value (or *fundamental* value) of the exchange rate are expected to be useful for predicting future changes of the exchange rate, especially at long-forecast horizons. This proposition is testable using long-horizon regression tests. In a landmark study, Mark (1995) provided evidence that current-period deviations from the equilibrium exchange rate help predict future changes in nominal exchange rates. Using data for the 1973.I–1991.IV period, Mark found a pattern of increased long-horizon predictability. While only some of his long-horizon regression test statistics were significant at conventional levels, Mark (1995, p. 215) conjectured that only the small sample size prevented more of his results from being significant.

The first contribution of this paper is to show that Mark's findings are not robust to extending the sample period up to 1997.IV. Contrary to Mark's conjecture, using his method of inference, there is less rather than more evidence of long-horizon predictability in the 1973.I–1997.IV period.[2] The second contribution of the paper is methodological. It is shown that the bootstrap procedure used by Mark is not entirely correct, and may result in spurious inference. After correcting for inconsistencies in the test procedure and for small-sample bias, the results of Mark's method can be reconciled with the results for the bootstrap method proposed in this paper. The third contribution of this paper is a point of interpretation. It is shown that the baseline results in Mark (1995) are open to misinterpretation and do not accurately measure the contribution of the monetary model to forecast performance. After suitably adjusting the test procedure, there is only very limited support for the monetary model and no evidence of increased long-horizon predictability. This finding is shown to be consistent with additional simulation evidence on the power of the long-horizon regression test. Finally, evidence is presented that the linear VEC model framework underlying the empirical study is likely to be misspecified, and that the methodology for constructing bootstrap p-values for long-horizon regression tests may be fundamentally flawed.

The remainder of the paper is organized as follows. Section 2 contains some useful statistical relationships based on the monetary exchange rate model which underlies the long-horizon regressions in Mark (1995). In Section 3, these relationships are used to construct a bootstrap test for long-horizon regressions. Section 4 presents the empirical findings. Section 5 analyses the size and power of the bootstrap test, and Section 6 contains the conclusions.

2. THE MONETARY MODEL IN VECTOR ERROR CORRECTION REPRESENTATION

In the standard long-run monetary model of exchange rate determination it is assumed that purchasing power parity and uncovered interest parity hold. Demand for log real balances is static and linearly related to log real income and the nominal interest rate. Denote the money demand income elasticity by λ and the money-demand interest rate semi-elasticity by ϕ. In the empirical part, λ will be set to 1 following Mark (1995). Further let $\delta \equiv \phi/(1 + \phi)$. In the absence

[2] As one referee pointed out, one plausible explanation of this result may be structural change in the economies of Germany and Japan in the 1990s. I do not pursue this explanation in this paper.

of speculative bubbles, the model implies that the log exchange rate for two identical countries is determined by:

$$e_t = (1-\delta)E_t\left(\sum_{j=0}^{\infty}\delta^j f_{t+j}\right) \tag{1}$$

where $f_t \equiv (m_t - m_t^*) - \lambda(y_t - y_t^*)$, m_t is the money stock in logs, y_t is real income in logs, and * denotes the foreign country. Subtracting f_t from both sides and rearranging yields:

$$e_t - f_t = E_t\left(\sum_{j=1}^{\infty}\delta^j \Delta f_{t+j}\right) \tag{2}$$

Provided that f_t is a serially correlated stationary process in first differences, equation (1) implies that $e_t \sim \mathrm{I}(1)$ and by equation (2) $e_t - f_t \sim \mathrm{I}(0)$. Thus, e_t and f_t are cointegrated with cointegrating vector $C' = [1, -1]$, and f_t may be interpreted as the long-run equilibrium value (or *fundamental* value) of the spot exchange rate. The implied joint time series process for e_t and f_t may be represented as a bivariate vector autoregression (VAR) for $x_t = (e_t, f_t)'$:

$$x_t = \nu + \Phi_1 x_{t-1} + \cdots + \Phi_p x_{t-p} + u_t \tag{3}$$

where u_t is assumed to be i.i.d. white noise with vector mean zero and nonsingular covariance matrix $\Sigma_u = \mathrm{E}(u_t u_t')$ and $\nu = [\nu_e \quad \nu_f]'$ is the intercept. Let $z_t \equiv e_t - f_t$ denote the deviation of the spot exchange rate from its fundamental value. As noted by Berkowitz and Giorgianni (1997), the VAR model (3) may be rewritten in vector error correction (VEC) form as:

$$\Delta x_t = \nu + \xi_0 x_{t-1} + \xi_1 \Delta x_{t-1} + \cdots + \xi_{p-1}\Delta x_{t-p+1} + u_t \tag{4}$$

where $\xi_0 = -HC'$ is a (2×2) matrix with rank $r = 1$ and $H = (h_1, h_2)'$ and $C = (c_1, c_2)'$ are (2×1) vectors. Given $C' = [1, -1]$ we can write:

$$\xi_0 x_{t-1} = -H(C' x_{t-1}) = -H(e_{t-1} - f_{t-1}) = -Hz_{t-1}$$

Substituting this expression into equation (4) we obtain the VEC model:

$$\begin{aligned} e_t &= \nu_e + e_{t-1} - h_1 z_{t-1} + \xi_1^{11}\Delta e_{t-1} + \xi_1^{12}\Delta f_{t-1} + \cdots + \xi_{p-1}^{11}\Delta e_{t-p+1} + \xi_{p-1}^{12}\Delta f_{t-p+1} + u_{1t} \\ f_t &= \nu_f + f_{t-1} - h_2 z_{t-1} + \xi_1^{21}\Delta e_{t-1} + \xi_1^{22}\Delta f_{t-1} + \cdots + \xi_{p-1}^{21}\Delta e_{t-p+1} + \xi_{p-1}^{22}\Delta f_{t-p+1} + u_{2t} \end{aligned} \tag{4'}$$

Subtracting the second from the first equation in equation (4′) provides a solution for z_t:

$$z_t = (\nu_e - \nu_f) + \rho z_{t-1} + \tilde{u}_t \tag{5}$$

where $\rho = 1 - h_1 + h_2$ and the remainder term $\tilde{u}_t$ will in general be serially correlated.

3. BOOTSTRAPPING LONG-HORIZON REGRESSION TESTS

Numerous econometric studies have found that the random walk model provides more accurate forecasts than other models of the exchange rate (e.g. Meese and Rogoff, 1983, 1988; Diebold and Nason, 1990). Thus, the random walk model is a natural benchmark in judging forecast performance. The monetary model of Section 2 suggests that regressions of the form:

$$e_{t+k} - e_t = a_k + b_k z_t + \varepsilon_{t+k} \quad k = 1, 4, 8, 12, 16 \tag{6}$$

will improve forecast accuracy relative to the random walk forecast:

$$e_{t+k} - e_t = d_k + \varepsilon_{t+k} \quad k = 1, 4, 8, 12, 16 \tag{7}$$

by exploiting the mean reversion of z_t. This conjecture can be tested as $H_0 : b_k = 0$ versus $H_1 : b_k < 0$ for a given forecast horizon k, or jointly for all forecast horizons as $H_0 : b_k = 0\ \forall k$ versus $H_1 : b_k < 0$ for some k. In essence, this test is a standard Granger non-causality test for z_t in model (6) based on the full sample. Alternatively, the out-of-sample prediction mean-squared error of models (6) and (7) based on a sequence of recursive forecasts may be evaluated using Theil's U-statistic or the DM statistic of Diebold and Mariano (1995). A formal test compares the null of equal forecast accuracy against the one-sided alternative that forecasts from model (6) are more accurate than those from model (7). It is well known that asymptotic critical values for these test statistics are severely biased in small samples. In order to mitigate these size distortions critical values may be calculated based on the bootstrap approximation of the finite sample distribution of the test statistic under the null hypothesis of no exchange rate predictability in the cointegrated model (4′) or some equivalent representation of the data-generating process. Unlike asymptotic critical values, bootstrap critical values based on the percentiles of the bootstrap distribution automatically adjust for the increase in the dispersion of the finite-sample distribution of the test statistic that occurs in near-spurious regressions as the sample size grows. As a result, bootstrap inference is immune from the near-spurious regression problem discussed in Berkowitz and Giorgianni (1997). However, special care must be taken to ensure the validity of the bootstrap model under the null.[3]

3.1. Bootstrapping Long-horizon Regression Tests under the Null Hypothesis

A valid bootstrap algorithm under the maintained assumption of cointegration may be readily constructed from representation (4′). Under the null hypothesis of no exchange rate predictability the bootstrap data-generating process is obtained by fitting the restricted VEC model

$$\begin{aligned} \Delta e_t &= v_e + u_{1t} \\ \Delta f_t &= v_f - h_2 z_{t-1} + \sum_{j=1}^{p-1} \xi_j^{21} \Delta e_{t-j} + \sum_{j=1}^{p-1} \xi_j^{22} \Delta f_{t-j} + u_{2t} \end{aligned} \tag{8}$$

subject to the constraint that $h_2 < 0$, where p has been determined under H_0 by a suitable lag order selection criterion such as the Akaike Information Criterion (AIC). The restricted model by

[3] For a recent review of the bootstrap testing methodology in time series models see Li and Maddala (1996).

construction has the same i.i.d. innovations as model (4′). Under the null hypothesis of no exchange rate predictability, it is known that $h_1 = 0$ which imposes the restriction $h_2 < 0$ for z_t to be I(0).[4] This condition must be imposed in estimation to ensure the stationarity of the bootstrap data-generating process (DGP) for z_t in small samples. Estimation of model (8) thus requires the use of constrained estimated generalized least squares (EGLS) with all coefficients but v_e set equal to zero in the first equation and $(-h_2)$ constrained to be positive in the second equation.[5] After estimating model (8), pseudo data may be generated under the assumption of i.i.d. innovations by drawing with replacement from the residuals and recursively generating a sequence of bootstrap data conditional on the estimated parameters.[6] A detailed description of the algorithm can be found in the Appendix. Additional restrictions on the bootstrap DGP may arise in special cases. For example, the null hypothesis that the exchange rate is known to follow a random walk without drift implies the restrictions $v_e = 0$ in model (8) and $d_k = 0$ in the forecast model (7).[7] However, such an assumption is tenuous at best, and may result in spurious inference. Under the less restrictive assumption that the exchange rate follows a random walk, possibly with drift, v_e and d_k must remain unrestricted.[8]

[4] In the absence of augmented terms in the VEC model, this condition is both necessary and sufficient for the stationarity of z_t under H_0. More generally, in the presence of augmented terms, it is necessary, but not sufficient. Lack of sufficiency can be demonstrated by counterexamples. To establish necessity it is useful to rewrite model (5) as an ARMA process. For this ARMA process to be stationary, the largest root of the AR polynomial must be in the stationary region. For example, if there is one augmented term in the VEC model, the AR polynomial of z_t will be of second order with coefficients $1 + h_2 + \xi_1^{22}$ and $-\xi_1^{22}$. Imposing stationarity, the requirement that $h_2 < 0$ follows immediately from standard results for second-order difference equations (see Hamilton, 1994, p. 18). For higher-order AR polynomials no analytic solutions for the roots of the autoregressive polynomial exist, but a grid search over the parameter space confirms that $h_2 < 0$ is a necessary condition for the stationarity of z_t.

[5] In practice, h_2 may be constrained to some negative number ε, where ε is arbitrarily close to zero, provided $\varepsilon \to 0$ as $T \to \infty$. Under the null hypothesis, this constraint will not be binding asymptotically, so the asymptotic validity of the bootstrap procedure is not affected, regardless of the precise value of the constraint. EGLS estimation was implemented using an adaptation of the algorithm described in Lütkepohl (1991, p. 168). The asymptotic validity of this bootstrap procedure follows from the standard assumptions in Bose (1988) after observing that the VEC model in (4) may be equivalently represented as a VAR in Δe_t and z_t. Under the null hypothesis, the restricted EGLS estimator asymptotically converges to the standard LS estimator considered by Bose. Note that the discontinuity in the asymptotic distribution discussed in Basawa *et al.* (1991) does not arise in this model, because the cointegrating vector has been imposed in the vector error correction model.

[6] The assumption of i.i.d. innovations is not controversial for the quarterly data used in this paper. However, time-varying volatility is frequently found in financial data (including exchange rates) sampled at monthly or weekly intervals. In the presence of higher-order dependence the bootstrap procedure discussed here must be suitably modified in one of two ways. If the functional form of the higher-order dependence is known or can be estimated, it may be imposed in modelling the error term. For example, Lamoureux and Lastrapes (1990) model persistence in the variance of stock returns as GARCH. Their bootstrap algorithm could be adapted easily to model the residuals of the restricted VEC model (8) as GARCH. The VEC-GARCH model can be interpreted as a special case of the well-known class of VAR-GARCH models in empirical finance (e.g. Bekaert *et al.*, 1997). Alternatively, the dependence of the residuals may be modelled non-parametrically by resampling blocks of residuals, based on results by Künsch (1989) and others. Li and Maddala (1997) recently applied this strategy to the residuals of cointegrating regression models, and a similar sitrategy would be appropriate for the residuals of model (8) in the presence of time-varying volatility of unknown form.

[7] Note that under the null hypothesis of a random walk without drift the intercept in model (6) will be zero. An intercept must be included, however, because under the alternative hypothesis z_t enters with possibly non-zero mean.

[8] For example, Diebold, Gardeazabal, and Yilmaz (1994, p. 732) argue for including a drift, unless there is irrefutable evidence to the contrary.

3.2. Comparison with Earlier Bootstrap Long-horizon Regression Tests

Given the null hypothesis of no exchange rate predictability, Mark (1995) postulates the bootstrap DGP:

$$\begin{aligned} e_t - e_{t-1} &= a_0 + \varepsilon_{1t} \\ z_t &= b_0 + \sum_{j=1}^{J} b_j z_{t-j} + \varepsilon_{2t} \end{aligned} \tag{9}$$

where the innovations $\varepsilon_t = (\varepsilon_{1t}, \varepsilon_{2t})'$ are i.i.d. and $\Sigma_\varepsilon = \mathrm{E}(\varepsilon_t \varepsilon_t')$. He estimates each equation of this model by OLS and generates bootstrap data conditional on the fitted values, possibly after correcting for bias in the second equation.[9]

A simple example will illustrate how this bootstrap procedure relates to the bootstrap procedure described in Section 3.1. Suppose that the exchange rate follows a random walk and e_t and f_t are cointegrated such that $z_t \sim \mathrm{I}(0)$. For simplicity further suppose that there is just one lagged difference in model (4′). Then under H_0:

$$\begin{aligned} e_t &= v_e + e_{t-1} + u_{1t} \\ f_t &= v_f + f_{t-1} - h_2(e_{t-1} - f_{t-1}) + \xi_1^{21}(e_{t-1} - e_{t-2}) + \xi_1^{22}(f_{t-1} - f_{t-2}) + u_{2t} \end{aligned} \tag{10}$$

This VEC model may be expressed as a subset VAR in Δe_t and z_t. Pre-multiplying model (10) by a comfortable identity matrix whose second row has been replaced by C' yields an equivalent representation based on Campbell and Shiller (1987):

$$\begin{aligned} \Delta e_t &= v_e + u_{1t} \\ z_t &= v_e - v_f + (1 + h_2 + \xi_1^{22}) z_{t-1} - \xi_1^{22} z_{t-2} - (\xi_1^{21} + \xi_1^{22}) \Delta e_{t-1} + u_{1t} - u_{2t} \end{aligned} \tag{10′}$$

By substituting for the lagged Δe_t in the second equation of (10′), one may express the system in terms of the two marginal time series processes for Δe_t and z_t:

$$\begin{aligned} \Delta e_t &= v_e + u_{1t} \\ z_t &= (1 + \xi_1^{21} + \xi_1^{22}) v_e - v_f + (1 + h_2 + \xi_1^{22}) z_{t-1} - \xi_1^{22} z_{t-2} + u_{1t} - (\xi_1^{21} + \xi_1^{22}) u_{1t-1} - u_{2t} \end{aligned} \tag{10″}$$

The second equation of this system is the sum of a white noise process u_{2t} and an ARMA(2,1) process in z_t and u_{1t}. Engel (1984) proves that the sum of two possibly correlated ARMA processes will remain an ARMA process. This suggests approximating the ARMA process for z_t in (10″) by a suitable higher-order AR process, which results in Mark's model (9). Provided that the estimated process for z_t is stationary and care is taken to include a sufficient number of autoregressive lags, the bootstrap critical values from model (9) will be asymptotically equivalent to those from model (8). However, in finite samples, they will be inefficient, if the model is estimated by equation-by-equation least-squares methods rather than constrained EGLS.

[9] In related work, Campbell (1993) considers a special case of the bootstrap algorithm in Mark (1995). In his model $J = p = 1$.

4. EMPIRICAL RESULTS

The data for this paper was constructed from *OECD Main Economic Indicators* and country source data for 1973.I–1997.IV. It includes the US dollar exchange rates of the Canadian dollar, the German mark, the Japanese yen, and the Swiss franc, and the corresponding fundamentals. All data have been transformed exactly as described in Mark (1995). The data source is Datastream.[10]

All test results will be presented in the form of bootstrap *p*-values based on 2000 bootstrap replications.[11] Figures 1, 2, 3, 5 and 6 show the bootstrap *p*-values for a number of key statistics. t_{20} and t_{A} are the *t*-statistics for the slope coefficient in the long-horizon regression, with the subscript indicating whether the robust standard error is calculated based on a fixed truncation lag of 20 or Andrews' (1991) procedure. DM_{20} and DM_{A} refer to the corresponding Diebold–Mariano statistics and *U* to Theil's *U*-statistic. Note that the test results for individual *t*-statistics must not be viewed as independent tests since $b_k = b_1 \sum_{i=0}^{k-1} \rho^i$ in model (6) (see Berkowitz and Giorgianni, 1997). To circumvent this problem, Mark suggests bootstrapping the distribution of the infimum of the *t*-statistics across the five time horizons of interest. In the figures, this statistic is labelled the joint *t*-statistic. In a similar manner, joint tests are constructed for the other statistics.

There are two main criteria for assessing exchange rate predictability. One criterion is whether the joint test statistic is significant at the 10% level (corresponding to a *p*-value below 0·100). The other criterion is evidence of declining *p*-values as the forecast horizon is increased. We begin by contrasting the results obtained by Mark (1995) for the original sample period (in Figure 1(a)) with the results for the updated sample (in Figure 1(b)).[12] For the original data set (1973.I–1991.IV), most joint statistics indicate overall predictability of the DM and yen exchange rate, and there is considerable evidence of declining *p*-values for all countries but Canada. This finding led Mark to conjecture that only the small sample size prevented a more complete vindication of the monetary model. With the benefit of hindsight, we are in a position to verify this conjecture. For the sample period 1973.I–1997.IV, in Figure 1(b) we find that only for Switzerland is there some evidence of overall predictability, as measured by the joint statistics. Moreover, virtually all *p*-values are stable or *increasing* as the forecast horizon is increased. Hence, contrary to the original findings, there appears to be no evidence of increased long-horizon predictability. Thus, the monetary model of exchange rate determination is rejected for three of the four countries in the sample.

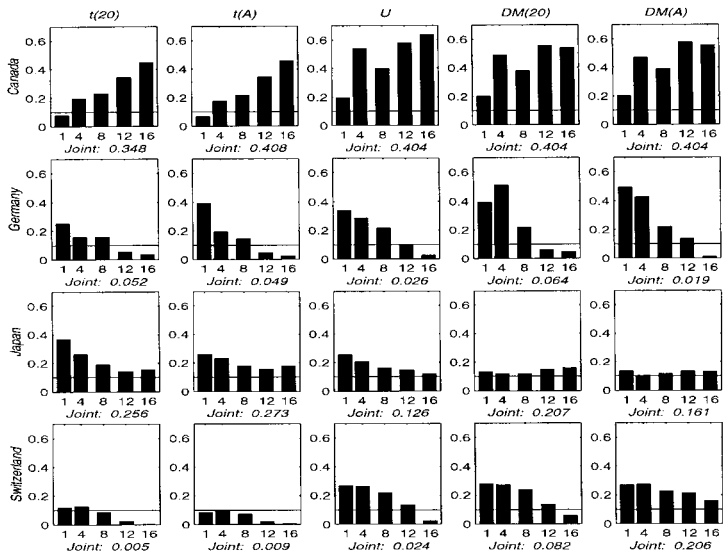
However, these results have to be interpreted with caution. One reason is that the bootstrap algorithm used in Mark (1995) is not entirely correct. In particular, note that Mark's procedure allows for a possible drift in the exchange rate in specifying the bootstrap replica of the

[10] Throughout this paper, the lag orders for the bootstrap DGPs will be selected using the AIC allowing for up to 8 lags for model (9) and up to 4 lags for model (8). Since the Jarque–Bera test rejects the null of Gaussian innovations for some countries, all bootstrap inference in this paper will be based on non-parametric resampling of the residuals.

[11] No slope coefficients are reported because, under the alternative hypothesis, the slope coefficients by construction will increase with the forecast horizon, so that evidence of increasing slopes does not imply increased long-horizon predictability (see Berkowitz and Giorgianni, 1997). This observation applies whether or not the slope coefficients are bias-adjusted. Similarly, statistical or visual measures of in-sample fit alone cannot be regarded as informative. Therefore, in this paper, only marginal significance levels are presented. The use of bootstrap *p*-values also avoids the problem of spurious fits discussed in Berkowitz and Giorgianni (1997).

[12] The lag order selection procedure for the extended data set differs slightly from Mark (1995) in that the lag order *J* is selected by the AIC, given an upper bound of 8 lags. The two methods give virtually identical results for the original Mark data set.

(a) 1973.I-1991.IV



(b) 1973.I-1997.IV

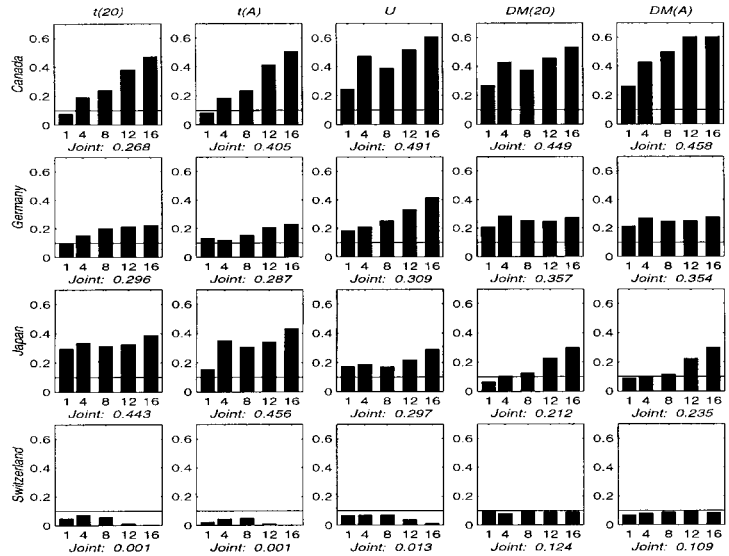


Figure 1. Bootstrap p-values from Mark's (1995) bootstrap model restricted under H_0: driftless random walk in the exchange rate. Here and in Figures 2, 3, 5 and 6 for a description of the statistics and models see text. Results are shown for alternative forecast horizons $k = 1, 4, 8, 12, 16$. Values below the horizontal line are significant at the 10% level. *Joint* refers to the p-value for the joint test statistic for all horizons

population process ($\nu_e \neq 0$), while ignoring this same drift in constructing the no-change forecast of the exchange rate ($d_k = 0$). Because the bootstrap model is not consistent with the model under H_0, the resulting bootstrap critical values will not be correct, and inference spurious. This mistake is akin to that of looking up the wrong Dickey–Fuller table for a unit root test. If we want to test the long-horizon regression model against the random walk model without drift, it is essential that we restrict the drift term in the bootstrap model to zero ($\nu_e = 0$). Figure 2(a) shows the results for the extended sample after imposing this consistency constraint. The effect is to lower the *p*-values of the out-of-sample statistics relative to Figure 1(b). For Switzerland, the evidence of overall predictability improves considerably. All five joint statistics are significant. For Canada and Japan the magnitude of the *p*-values of the joint test is cut in half. There is little effect on the pattern of *p*-values across forecast horizons.

A second reason for caution is the possible presence of small-sample bias in the ordinary least-squares estimator. To the extent that there is bias in the coefficient estimates, the estimated bootstrap DGP will not be representative for the underlying population process, and the resulting bootstrap critical values will be misleading (see Kilian, 1998). Mark (1995, p. 208) investigated the effect of small-sample bias for the original data set and concluded that his results were not sensitive. We now investigate the same question for the extended data set.[13]

The dominant roots ρ of the estimated processes in Figure 1(b) are 0·9166, 0·9054, 0·9188, and 0·8827, respectively. Using the bias corrections of Shaman and Stine (1988) the bias-corrected roots are 0·9254, 0·9140, 0·9761, and 0·8889. Thus, there appears to be strong bias in the bootstrap DGP for Japan, but not much bias in the other DGPs. Figure 2(b) reports the effects of correcting for small-sample bias in the bootstrap DGP (without imposing the consistency constraint). As expected, the differences in *p*-values are comparatively minor with the exception of Japan. Compared to Figure 1(b), the evidence of overall predictability for Japan improves considerably. Two of the three out-of-sample test statistics become jointly significant and the *p*-value of the third one drops from 0·297 to 0·137. This example shows that small-sample bias-corrections can be important in practice.

It would be a mistake to consider the two corrections presented in Figure 2 in isolation, because they tend to interact. Figure 3(a) therefore presents results for the extended data set that incorporate the consistency constraint as well as bias corrections for the bootstrap DGP. The net result is strong overall predictability for Switzerland (with all joint statistics significant), considerable evidence of out-of-sample predictability for Japan, and a sizable reduction in the magnitude of the *p*-values of the out-of-sample statistics for Canada. These results confirm that the bootstrap methodology matters for the interpretation of the results.

Apart from the aforementioned internal inconsistency in Mark's (1995) bootstrap procedure and small-sample bias, the main remaining difference between his bootstrap method and the VEC bootstrap procedure proposed in this paper are differences in lag order selection and in estimation efficiency. Figure 3(b) shows the corresponding results for the VEC bootstrap. With the lag order constrained to lie between 0 and 4, the AIC selects four augmented lags for Canada, zero for Germany, and two each for Japan and Switzerland. The constraint on h_2 is binding only for Germany. The implied roots for the other z_t processes are 0·9344, 0·9925, and 0·9340, respectively. They tend to be higher than the bias-corrected roots for the Mark procedure.

[13] One potentially important difference between Mark's study and this study is the method of bias correction. Mark used an *ad hoc* bias adjustment without a firm basis in statistical theory. In contrast, the bias estimates used in this paper are based on the closed form solutions derived in Shaman and Stine (1988).

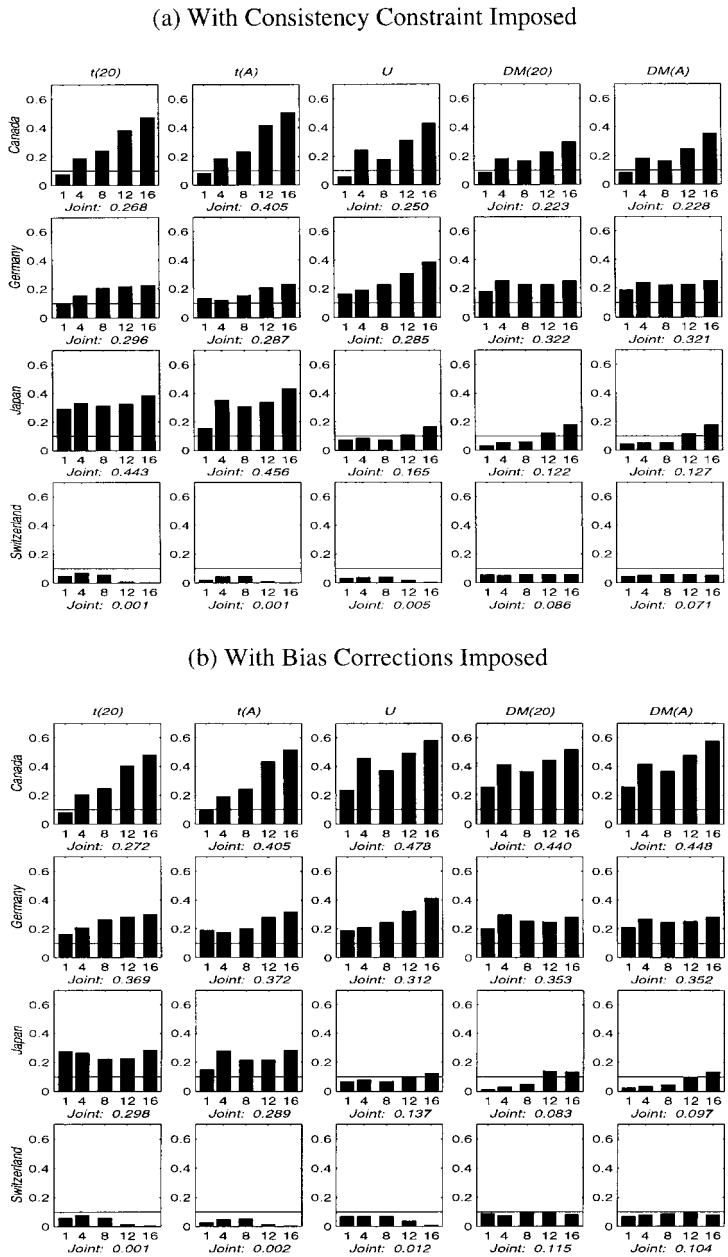
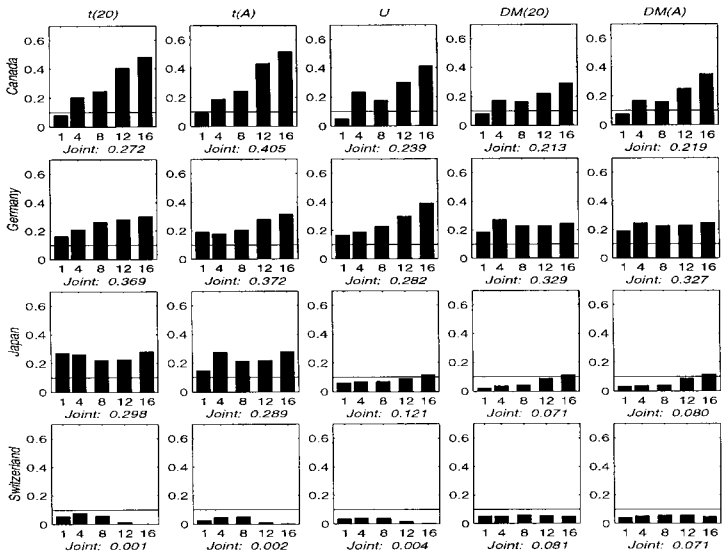


Figure 2. Bootstrap p-values from Mark's (1995) bootstrap model restricted under H_0: driftless random walk in the exchange rate

(a) Mark's (1995) Bootstrap Model With Bias Corrections and Consistency Constraint Imposed



(b) VEC Bootstrap Model

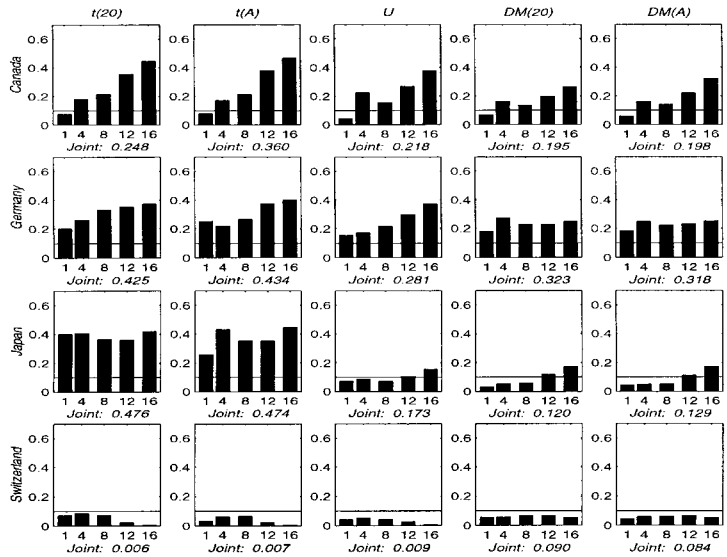


Figure 3. Bootstrap p-values under H_0: driftless random walk in the exchange rate

As Figure 3(b) shows, the remaining differences appear to be of minor consequence, with the exception of Japan, where the VEC model bootstrap detects no overall predictability in contrast to the results in Figure 3(a). However, the differences in the actual *p*-values are not large. Overall, the results in Figures 3(a) and 3(b) are remarkably similar. This fact lends additional credibility to the results. In essence, we find that the monetary model does not beat the random walk for at least two of the four currencies, and there is no evidence of increased long-horizon predictability.

An additional plausibility check of these results is provided by the calculation of persistence profiles. Pesaran and Shin (1996) propose measuring the strength of the mean reversion of the error correction term by calculating persistence profiles that measure the response of the error correction term z_t to a shock drawn from the multivariate distribution of u_t in model (4). These persistence profiles measure the speed of convergence of the exchange rate to its long-run equilibrium value following a disturbance of the equilibrium.[14] Figure 4 shows that for Switzerland the adjustment is complete 6 years after the shock. For Canada, equilibrium is restored after about 10 years. For Japan and Germany, even after 10 years the exchange rate has not returned to its equilibrium value, suggesting that deviations from equilibrium are highly persistent. Thus, one would expect predictability to exist, if at all, for the Swiss franc, and possibly for the Canadian dollar, but not for the Japanese yen or the DM.

The persistence profiles are roughly consistent with the pattern of the *p*-values for the in-sample *t*-statistics, but not with the out-of-sample results. In particular, the comparatively low out-of-sample *p*-values for Japan and the comparatively high out-of-sample *p*-values for Canada in Figure 3 may appear puzzling. This puzzle may be resolved by keeping in mind that low out-of-sample *p*-values in Figure 3 do *not* establish that economic fundamentals are responsible for the improved forecast accuracy; rather they measure the *joint* contribution of the drift term *and* the error correction term in the long-horizon regression forecast. Note that the out-of-sample statistics used in Mark (1995) and in Figures 1–3 compare the long-horizon regression forecast (equation 6) with the forecast based on the driftless random walk.

$$e_{t+k} - e_t = \varepsilon_{t+k} \quad k = 1, 4, 8, 12, 16 \tag{7'}$$

Thus, the superior out-of-sample accuracy of regression (6) may be due to the fact that this regression picks up an apparent drift in the exchange rate over the sample period or due to the inclusion of the error correction term. The reason for the improved forecast performance is not identified. This makes it impossible to interpret a significant improvement in forecast accuracy as evidence in favour of monetary exchange rate models. The out-of-sample statistics may either overstate or understate the true contribution of the fundamental by lumping its effect together with that of the drift term.

To isolate the *marginal* contribution of z_t, one must allow for a drift in the random walk forecast in model (7) as well as in the bootstrap DGP. The results are displayed in Figures 5(a) and 5(b). The inclusion of a drift does not affect the overall results for the in-sample statistics, but it leads to a striking change in the out-of-sample statistics. Both bootstrap procedures now detect important evidence of overall predictability for Canada as well as Switzerland. The evidence of overall predictability for Japan vanishes. For Germany, the changes are inconsequential. The

[14] Note that this exercise is fundamentally different from testing for cointegration. Here we are concerned not with the *existence* of cointegration, but with the speed of mean reversion of the error correction term as a measure of the *strength* of the cointegrating relationship.

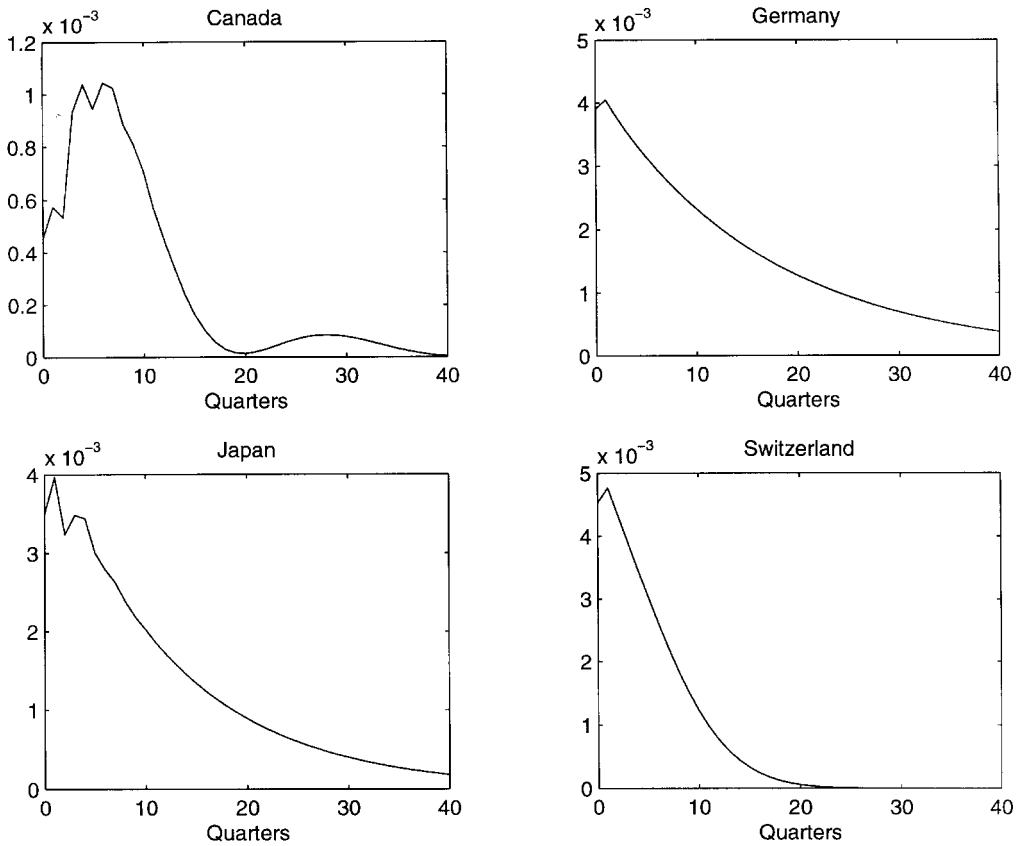


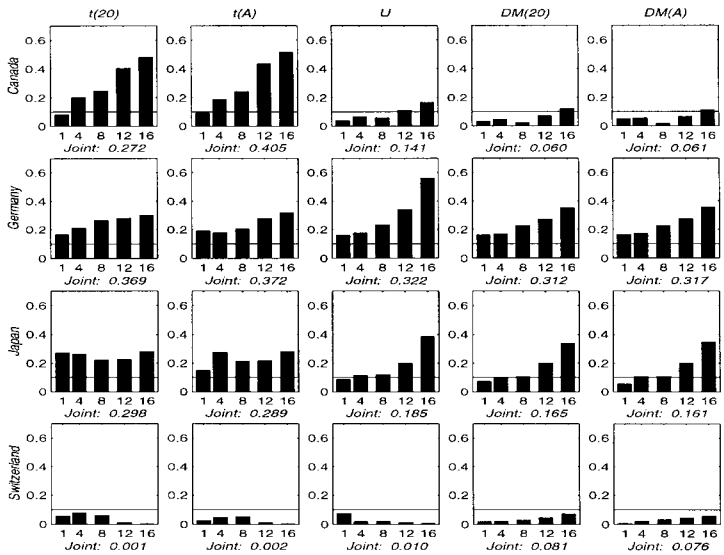
Figure 4. Speed of convergence of the exchange rate to its long-run equilibrium value. The persistence profiles were estimated using the methodology of Pesaran and Shin (1996). The plots show the response of z_t to a composite shock drawn from the multivariate error distribution of the VEC model. The results are for the extended data set for 1973.I–1997.IV. Lag orders were selected using the AIC and allowing for up to four augmented lags in the VEC representation

pattern of *p*-values across horizons remains flat or increasing in virtually all cases, indicating the absence of increased long-horizon predictability.

The results in Figure 5 suggest that monetary variables do help improve forecast accuracy relative to the random walk forecast at best for some countries, but not for all. The evidence of predictability in Figure 5 is actually stronger than the evidence in Figure 1 obtained by using Mark's original procedure, but it is weaker than what Mark (1995) conjectured it would be after extending the sample. How do we interpret these mixed results?

First, given that the model should work equally well for all countries, if it is an adequate representation of reality, the results cast doubt on the monetary model of exchange rate determination, at least in the simple form presented in Section 2. In this context, it would be worthwhile to investigate in future research whether or not the failure of the monetary model for

(a) Mark (1995) bootstrap model with bias corrections and consistency constraint



(b) VEC bootstrap model

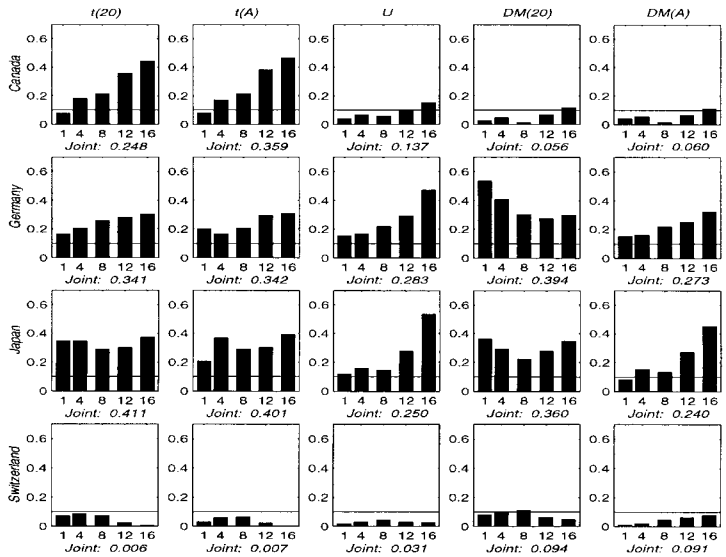


Figure 5. Bootstrap p-values under H_0: random walk with drift in the exchange rate

Japan and Germany is related to structural changes in those economies in the 1990s, namely unification in Germany and the financial crisis in Japan.

Second, and more importantly, the results suggest that more often than not, the monetary model is easier to reject at longer horizons. This finding is puzzling, given the widespread presumption that the power of long-horizon regression tests improves at longer horizons. The idea of increasing long-horizon predictability seems hard to reconcile with the fairly stable or increasing pattern of *p*-values in Figure 5. To investigate this puzzle it is useful to examine the size and power of the bootstrap test for several DGPs calibrated to the actual data.

5. THE SIZE AND POWER OF LONG-HORIZON REGRESSION TESTS

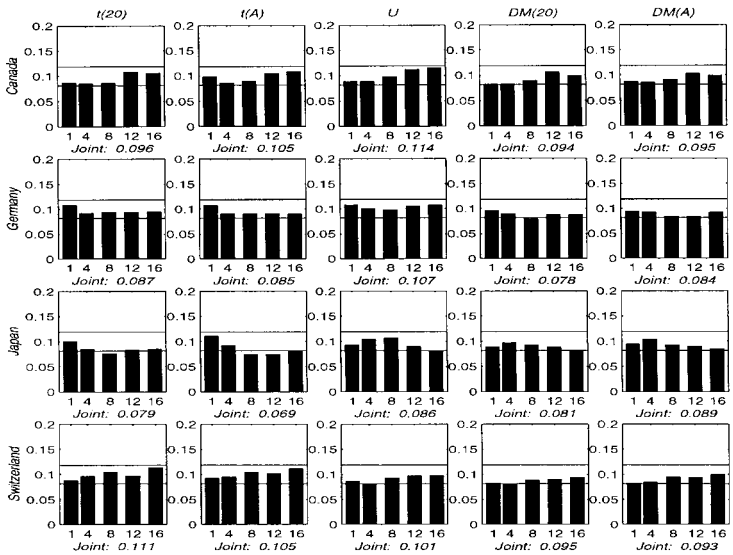
Figure 6 shows the effective size of the nominal 10% test based on bootstrap *p*-values. All results are based on 1000 trials with 2000 bootstrap replications each. The Monte Carlo standard error for a test at the 10% level is 0·0095. The DGP is based on the restricted VEC model under the null hypothesis that the exchange rate follows a random walk (possibly with drift) and the exchange rate and the fundamental are cointegrated. Separate DGPs are estimated for each country. The lag orders are based on the AIC as in all previous applications. The sample size is the same as for the extended data set. For each trial, the VEC bootstrap procedure of Figure 5(b) is used to calculate the *p*-values, and the rejection rates at the 10% level are tabulated. Figure 6(a) shows that, with few exceptions, the bootstrap test is remarkably accurate, considering the small sample size. Moreover, the size is roughly constant across forecast horizons. Overall, there is strong evidence that any systematic differences between test results for short and long horizon tests must be due to differences in power.

Figure 6(b) shows the power of the nominal 10% test against the alternative of an unrestricted VEC model. This paricular class of DGPs is the natural alternative to consider because it is implied by the theoretical model we are interested in testing. The DGPs are based on the best-fitting unrestricted VEC model estimates for Canada, Germany, Japan, and Switzerland. The simulation results suggest that there are no power advantages to long-horizon regression tests. Power tends to be constant across forecast horizons or declining. In the few cases in which power increases with the forecast horizon, the increase is neither statistically nor economically significant.[15]

The apparent absence of increasing long-horizon power against the VEC model alternative does not rule out that long-horizon tests have higher power against other alternatives. For example, it is often suggested that long-run predictability of the exchange rate may be the result of non-linear dynamics caused by regime switching or peso problems. It is easy to see how in non-linear models the power of long-horizon regression tests may be higher at longer horizons. Thus, it may be tempting to brush aside the evidence of constant or declining power for the linear VEC model as irrelevant given those alternative rationales for higher power. However, it is important to keep in mind that bootstrap *p*-values calculated under the maintained assumption of a linear VEC model are by construction invalid, if the true process is non-linear. Thus, we *cannot*

[15] The power results for the *t*-statistics are consistent with findings in Campbell (1993) for a simpler model. They also are consistent with Monte Carlo evidence for the exact finite sample distributions of the test statistic in Bollerslev and Hodrick (1995, p. 434). In related work, Berben and van Dijk (1998) prove that the local asymptotic power of the *t*-test is independent of the forecast horizon in a root-local-to-unity setting. No analytic results exist for the out-of-sample statistics.

(a) Size of Nominal 10 Percent Test



(b) Power against VEC model alternative

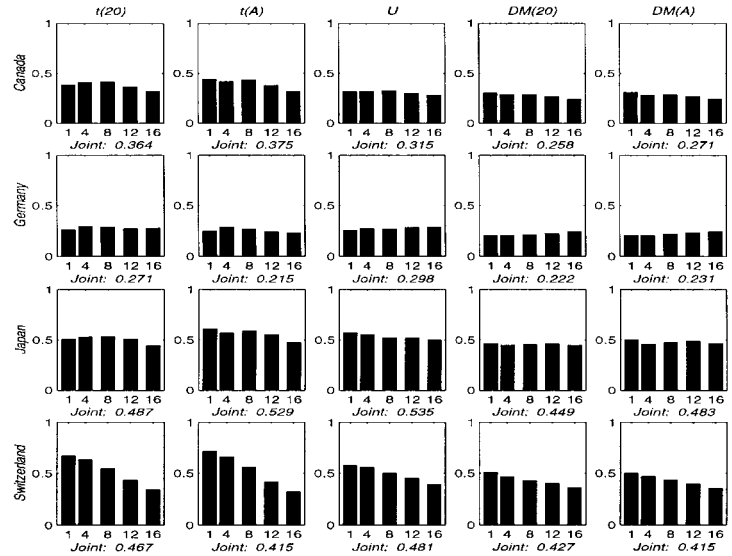


Figure 6. Size and power of VEC model bootstrap test

interpret evidence in favour of predictability based on such p-values as support for explanations based on non-linear mean reversion in z_t. To do so, would require a suitably modified procedure for bootstrapping long-horizon regression tests under the maintained assumption of a specific non-linear model. The development of such a procedure is the subject of ongoing research.

The case for abandoning the construction of bootstrap p-values based on the linear VEC model is strong. Given the stable rejection rates of the test under the null hypothesis, the observed pattern of p-values in Figure 5 for the actual data simply is not consistent with the rejection rates in Figure 6(b) under the alternative hypothesis. This observation is compelling evidence that the bootstrap DGP (and by implication the monetary model of Section 2) is likely to be misspecified, and it provides indirect support for the evidence of nonlinear mean reversion in z_t presented in Taylor and Peel (1998).

6. CONCLUSIONS

Numerous studies have documented severe size distortions of long-horizon regression tests. In this paper, a new bootstrap method for small-sample inference in long-horizon regressions was introduced. Simulation evidence confirmed that this bootstrap method greatly reduces the size distortions of conventional long-horizon regression tests in small samples.

The use of this bootstrap procedure was illustrated by analysing the long-horizon predictability of four exchange rates under the current float, and the findings were reconciled with those of an earlier study by Mark (1995). There was some evidence of exchange rate predictability, but, contrary to earlier studies, no evidence of higher predictability at longer forecast horizons. The latter finding is consistent with the results of a Monte Carlo study that long-horizon regression tests tend to have stable or declining power against the unrestricted VEC model alternative implied by the monetary theory of exchange rate determination.

Perhaps the most interesting finding to emerge from this study is that the linear VEC model framework underlying the existing long-horizon regression tests is likely to be misspecified. In particular, the observed pattern of p-values across forecast horizons in the empirical study is inconsistent with the size and power results for the linear VEC model. This fact is suggestive of a non-linear data-generating process (DGP).

Evidence of non-linearities may seem to vindicate previous findings of long-horizon predictability, but it is important to keep in mind that standard bootstrap p-values for long-horizon regression tests are invalid in the presence of non-linearities in the DGP. This does not mean that long-horizon regression tests should be abandoned, but it means that the construction of appropriate bootstrap p-values must be rethought. The presence of non-linearities in the DGP requires a new class of inferential procedures for long-horizon regression tests. The development of such procedures is the subject of ongoing research. Evidence from these procedures may fundamentally alter our views of the predictability of asset returns in general and of exchange rates in particular.

APPENDIX: BOOTSTRAP ALGORITHM FOR LONG-HORIZON REGRESSION TEST

The bootstrap algorithm used to generate the results in Figure 5 consists of four steps:

(1) Given the sequence of observations $\{x_t\}$, where $x_t = (e_t, f_t)'$, estimate the long-horizon regression:

$$e_{t+k} - e_t = a_k + b_k z_t + \varepsilon_{t+k} \quad k = 1, 4, 8, 12, 16$$

and construct the test statistic of interest, $\hat{\theta}$.

(2) The DGP for $x_t = (e_t, f_t)'$ is:

$$e_t = v_e + e_{t-1} - h_1 z_{t-1} + \xi_1^{11}\Delta e_{t-1} + \xi_1^{12}\Delta f_{t-1} + \cdots + \xi_{p-1}^{11}\Delta e_{t-p+1} + \xi_{p-1}^{12}\Delta f_{t-p+1} + u_{1t}$$
$$f_t = v_f + f_{t-1} - h_2 z_{t-1} + \xi_1^{21}\Delta e_{t-1} + \xi_1^{22}\Delta f_{t-1} + \cdots + \xi_{p-1}^{21}\Delta e_{t-p+1} + \xi_{p-1}^{22}\Delta f_{t-p+1} + u_{2t}$$

where the cointegration constraint implied by the net present value model has been imposed. The innovation term is assumed to be i.i.d. distributed. Estimate this model by EGLS subject to the constraints that $h_2 < 0$ and that all coefficients but v_e in the first equation are zero. It is assumed that the lag order p has been determined under H_0 by a suitable lag order selection criterion such as the AIC.

(3) Based on the fitted model generate a sequence of pseudo observations $\{x_t^*\}$ of the same length as the original data series $\{x_t\}$, where $x_t^* = (e_t^*, f_t^*)'$ is based on cumulative sums of the realizations of the bootstrap data-generating process:

$$\Delta e_t^* = \hat{v}_e + u_{1t}^*$$
$$\Delta f_t^* = \hat{v}_f - \hat{h}_2 z_{t-1}^* + \sum_{j=1}^{p-1} \hat{\xi}_j^{21}\Delta e_{t-j}^* + \sum_{j=1}^{p-1} \hat{\xi}_j^{22}\Delta f_{t-j}^* + u_{2t}^*$$

To initialize this process specify $z_{t-1}^* = 0$ and $\Delta x_{t-j}^* = (0, 0)'$ for $j = p-1, \ldots, 1$ and discard the first 500 transients. The pseudo innovation term $u_t^* = (u_{1t}^*, u_{2t}^*)'$ is random and drawn with replacement from the set of observed residuals $\hat{u}_t = (\hat{u}_{1t}, \hat{u}_{2t})'$. Repeat this step 2000 times.

(4) For each of the 2000 bootstrap replications $\{x_t^*\}$ estimate the long-horizon regression

$$e_{t+k}^* - e_t^* = a_k^* + b_k^* z_t^* + \varepsilon_{t+k}^* \quad k = 1, 4, 8, 12, 16$$

and construct the test statistics of interest, $\hat{\theta}^*$.

(5) Use the empirical distribution of the 2000 replications of the bootstrap test statistic $\hat{\theta}^*$ to determine the p-value of the test statistic $\hat{\theta}$.

ACKNOWLEDGEMENTS

The comments of Bob Barsky, Jeremy Berkowitz, Ufuk Demiroglu, Frank Diebold, Lucia Fedina, Phil Howrey, Lorenzo Giorgianni, Jan Kmenta, Nelson Mark, Andrew Rose, Shinichi Sakata, Peter Schmidt, Matthew Shapiro, and Mark Taylor helped improve the paper. I also thank Steven Durlauf and two anonymous referees for their advice. Jeremy Berkowitz, Lorenzo Giorgianni, and Nelson Mark provided ready access to their data and programs.

REFERENCES

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

Basawa, I. V., A. K. Mallik, W. P. Cormick, J. H. Reeves and R. L. Taylor (1991), 'Bootstrapping unstable first-order autoregressive processes', *Annals of Statistics*, **19**, 1098–1101.

Bekaert, G., R. J. Hodrick and D. A. Marshall (1997), 'On biases in tests of the expectations hypothesis of the term structure of interest rates', *Journal of Financial Economics*, **44**, 309–348.

Berben, R.-P. and D. van Dijk (1998), 'Does the absence of cointegration explain the typical findings in long-horizon regressions?', mimeo, Tinbergen Institute, Erasmus University Rotterdam.

Berkowitz, J. and L. Giorgianni (1997), 'Long-horizon exchange rate predictability?' Working Paper No. 97/6, International Monetary Fund.

Bollerslev, T. and R. J. Hodrick (1995), 'Financial market efficiency tests', in M. H. Pesaran and M. R. Wickens (eds), *Handbook of Applied Econometrics*, Blackwell, Cambridge, MA.

Bose, A. (1988), 'Edgeworth correction by bootstrap in autoregressions', *Annals of Statistics*, **16**, 1709–1722.

Campbell, J. Y. (1993), 'Why long-horizons: a study of power against persistent alternatives'. NBER Technical Working Paper No. 142.

Campbell, J. Y. and R. J. Shiller (1987), 'Cointegration and tests of present value models', *Journal of Political Economy*, **95**, 1062–1088.

Campbell, J. Y. and R. J. Shiller (1988), 'Stock prices, earnings, and expected dividends', *Journal of Finance*, **43**, 661–676.

Chinn, M. D. and R. A. Meese (1995), 'Banking on currency forecasts: how predictable is change in money?' *Journal of International Economics*, **38**, 161–178.

Cutler, D. M., J. M. Poterba and L. H. Summers (1991), 'Speculative dynamics', *Review of Economic Studies*, **58**, 529–546.

Diebold, F. X. and R. Mariano (1995), 'Comparing predictive accuracy', *Journal of Business and Economic Statistics*, **13**, 253–262.

Diebold, F. X., J. Gardeazabal and K. Yilmaz (1994), 'On cointegration and exchange rate dynamics', *Journal of Finance*, **49**, 727–735.

Diebold, F. X. and J. A. Nason (1990), 'Nonparametric exchange rate prediction?', *Journal of International Economics*, **28**, 315–332.

Engel, E. M. R. A (1984), 'A unified approach to the study of sums, products, time-aggregation and other functions of ARMA processes', *Journal of Time Series Analysis*, **5**, 159–171.

Fama, E. F. and R. R. Bliss (1987), 'The information in long-maturity forward rates', *American Economic Review*, **77**, 680–692.

Fama, E. F. and K. R. French (1988), 'Dividend yields and expected stock returns', *Journal of Financial Economics*, **22**, 3–25.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton University Press, Princeton, NJ.

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

Kilian, L. (1998), 'Small-sample confidence intervals for impulse response functions', *Review of Economics and Statistics*, **80**, 218–230.

Kirby, C. (1997), 'Measuring the predictable variation in stock and bond returns', *Review of Financial Studies*, **10**, 579–630.

Künsch, H. R. (1989), 'The jackknife and the bootstrap for general stationary observations', *Annals of Statistics*, **17**, 1217–1241.

Lamoureux, C. G. and W. D. Lastrapes (1990), 'Persistence in variance, structural change, and the GARCH model', *Journal of Business and Economic Statistics*, **8**, 225–234.

Li, H. and G. S. Maddala (1997), 'Bootstrapping cointegrating regressions', *Journal of Econometrics*, **80**, 297–318.

Li, H. and G. S. Maddala (1996), 'Bootstrapping time series models', *Econometric Reviews*, **15**, 115–158.

Lütkepohl, H. (1991), *An Introduction to Multiple Time Series Analysis*, Springer-Verlag, New York.

Mankiw, N. G., D. Romer and M. D. Shapiro (1991), 'Stock market forecastability and volatility: a statistical appraisal', *Review of Economic Studies*, **58**, 455–477.

Mark, N. C. (1995), 'Exchange rates and fundamentals: evidence on long-horizon predictability', *American Economic Review*, **85**, 201–218.
Meese, R. and K. Rogoff (1983), 'Empirical exchange rate models of the 1970s: do they fit out of sample?' *Journal of International Economics*, **14**, 3–24.
Meese, R. and K. Rogoff (1988), 'Was it real? The exchange-rate interest differential relation over the modern floating-rate period', *Journal of Finance*, **4**, 933–948.
Nelson, C. R. and M. J. Kim (1993), 'Predictable stock returns: the role of small-sample bias', *Journal of Finance*, **48**, 641–661.
Pesaran, M. H. and Y. Shin (1996), 'Cointegration and speed of convergence to equilibrium', *Journal of Econometrics*, **71**, 117–143.
Shaman, P. and R. A. Stine (1988), 'The bias of autoregressive coefficient estimators', *Journal of the American Statistical Association*, **83**, 842–848.
Taylor, M. P. and D. A. Peel (1998), 'Non-linear adjustment, long-run equilibrium and exchange rate fundamentals', mimeo, Department of Economics, Oxford University.

