TESTS OF EQUAL FORECAST ACCURACY AND ENCOMPASSING FOR NESTED MODELS

Todd E. Clark
Michael W. McCracken

OCTOBER 1999

RWP 99-11

LAST REVISED: NOVEMBER 2000

Research Division
Federal Reserve Bank of Kansas City

Todd E. Clark is an assistant vice president and economist at the Federal Reserve Bank of Kansas City. Michael W. McCracken is an assistant professor of economics at Louisiana State University. The authors gratefully acknowledge the helpful comments of: Ken West, anonymous referees, Charles Engel, Lutz Kilian, Norm Swanson, and Dek Terrell; seminar participants at UCSD, the University of Kansas, the University of Michigan, and Penn State University; and participants at the 2000 World Congress of the Econometric Society and the 2000 NBER Summer Institute seminar on forecasting and time series methods. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Kansas City or the Federal Reserve System.

Clark e-mail: todd.e.clark@kc.frb.org
McCracken e-mail: mmccrac@unix1.sncc.lsu.edu

Abstract

We examine the asymptotic and finite-sample properties of tests for equal forecast accuracy and encompassing applied to 1-step ahead forecasts from nested linear models. We first derive the asymptotic distributions of two standard tests and one new test of encompassing and provide tables of asymptotically valid critical values. Monte Carlo methods are then used to evaluate the size and power of tests of equal forecast accuracy and encompassing. The simulations indicate that post-sample tests can be reasonably well sized. Of the post-sample tests considered, the encompassing test proposed in this paper is the most powerful. We conclude with an empirical application regarding the predictive content of unemployment for inflation.

Keywords: causality, forecast accuracy, forecast encompassing

JEL Nos.: C53, C12, C52

Abstract

We examine the asymptotic and finite-sample properties of tests for equal forecast accuracy and encompassing applied to 1-step ahead forecasts from nested linear models. We first derive the asymptotic distributions of two standard tests and one new test of encompassing and provide tables of asymptotically valid critical values. Monte Carlo methods are then used to evaluate the size and power of tests of equal forecast accuracy and encompassing. The simulations indicate that post-sample tests can be reasonably well sized. Of the post-sample tests considered, the encompassing test proposed in this paper is the most powerful. We conclude with an empirical application regarding the predictive content of unemployment for inflation.

Keywords: causality, forecast accuracy, forecast encompassing

JEL Nos.: C53, C12, C52

Tests of Equal Forecast Accuracy and Encompassing for Nested Models

Todd E. Clark and Michael W. McCracken*
Federal Reserve Bank of Kansas City and Louisiana State University

November 2000

Abstract

We examine the asymptotic and finite-sample properties of tests for equal forecast accuracy and encompassing applied to 1-step ahead forecasts from nested linear models. We first derive the asymptotic distributions of two standard tests and one new test of encompassing and provide tables of asymptotically valid critical values. Monte Carlo methods are then used to evaluate the size and power of tests of equal forecast accuracy and encompassing. The simulations indicate that post-sample tests can be reasonably well sized. Of the post-sample tests considered, the encompassing test proposed in this paper is the most powerful. We conclude with an empirical application regarding the predictive content of unemployment for inflation.

Keywords: causality, forecast accuracy, forecast encompassing

JEL Nos.: C53, C12, C52

**Clark (corresponding author)*: Economic Research Dept., Federal Reserve Bank of Kansas City, 925 Grand Blvd., Kansas City, MO 64198, todd.e.clark@kc.frb.org. *McCracken*: Dept. of Economics, Louisiana State University, 2107 CEBA Building, Baton Rouge, LA 70803, mmccrac@unix1.sncc.lsu.edu. We gratefully acknowledge the helpful comments of: Ken West, anonymous referees, Charles Engel, Lutz Kilian, Norm Swanson, and Dek Terrell; seminar participants at UCSD, the University of Kansas, the University of Michigan, and Penn State University; and participants at the 2000 World Congress of the Econometric Society and the 2000 NBER Summer Institute seminar on forecasting and time series methods. The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Kansas City or the Federal Reserve System.

1. Introduction

Since the influential work of Meese and Rogoff (1983, 1988), it has become common to use comparisons of out-of-sample forecasts to determine whether one variable has predictive power for another.[1] Typically, this out-of-sample comparison is made in two stages. First, forecasts of the variable of interest are constructed once using a model that includes a variable with putative predictive content and then a second time excluding that variable. Second, given the two sequences of forecast errors, tests of equal forecast accuracy or forecast encompassing are conducted. This out-of-sample approach is explicitly advocated by Ashley, Granger, and Schmalensee (1980), who argue that it is more in the spirit of the definition of Granger causality to employ post-sample forecast tests than to employ the standard full-sample causality test.

Although post-sample tests of this type are increasingly used, little is known about their effectiveness. Most evidence on the asymptotic and finite-sample behavior of tests of equal forecast accuracy and encompassing pertain to forecasts from non-nested models. Diebold and Mariano (1995), West (1996, 2000a, b), Harvey, Leybourne, and Newbold (1997, 1998), West and McCracken (1998), Clark (1999), Corradi, Swanson, and Olivetti (1999), and McCracken (2000) each present results for non-nested forecasts.

With nested models, however, test properties are likely to differ because, under the null, the forecast errors are asymptotically the same and therefore perfectly correlated. Only two extant studies focus on results for nested models. McCracken (1999) derives the asymptotic distributions of several tests of equal forecast accuracy between two nested models. Chao, Corradi and Swanson (2000) develop an out-of-sample test of causality that resembles an encompassing test applied to forecasts from nested models.

In this paper we first derive the limiting distributions of tests for encompassing applied to 1-step ahead forecasts from nested linear models. The encompassing tests are those proposed by Ericsson (1992) and Harvey, Leybourne and Newbold (1998) and a new statistic developed in this paper. As in West (1996, 2000a, b), West and McCracken (1998), Chao, Corradi, and Swanson (2000), Corradi, Swanson, and Olivetti (1999), and McCracken (2000), the limiting distributions explicitly account for the uncertainty introduced by parameter estimation.

In our results, when the number of observations used to generate initial estimates of the models and the number of forecast observations increase at the same rate, the limiting distributions of the tests are non-standard. We provide numerically-generated critical values for these distributions. However, when the number of forecasts increases at a slower rate than the number of observations used in the initial model estimates, the Ericsson (1992) and Harvey, Leybourne, and Newbold (1998) statistics are limiting standard normal.

We then use Monte Carlo simulations to examine the finite-sample size and size-adjusted power of the encompassing tests, as well as a set of equal mean square error (MSE) tests. These Monte Carlo experiments show that, in most settings, each of the post-sample tests is reasonably well sized when the statistics are compared against the asymptotic critical values provided in this paper and in McCracken (1999). However, comparing the post-sample forecast statistics against standard normal critical values usually makes the tests undersized. The Monte Carlo simulations also show that the powers of the tests permit some simple rankings, in which the new encompassing statistic proposed in this paper is most powerful in small samples.

Finally, to illustrate how the tests perform in practical settings, each test is used to determine whether the unemployment rate has predictive content for inflation in quarterly U.S.

[1] Examples of studies using this methodology include Diebold and Rudebusch (1991), Amano and van Norden (1995), Chinn and Meese (1995), Mark (1995), Krueger and Kuttner (1996), Bram and Ludvigson (1998),

data. We find the evidence mixed, but suggestive of a relationship. While each of the equal MSE tests fail to reject the null that unemployment has no predictive content for inflation, each of the encompassing tests indicates that unemployment does have predictive power.

Section 2 introduces the notation, the forecasting and testing setup, and the assumptions. Section 3 defines the forecast encompassing tests considered and provides the null asymptotic results. In the interest of brevity, proofs are provided in Clark and McCracken (2000). In Section 4 we present a Monte Carlo evaluation of the finite-sample size and power of tests of forecast encompassing and equal MSE. Section 5 uses the tests to determine whether the unemployment rate has predictive power for inflation.

2. Setup

The sample of observations $\{y_t, x'_{2,t}\}_{t=1}^{T+1}$ includes a scalar random variable y_t to be predicted and a ($k_1 + k_2 = k\times1$) vector of predictors $x_{2,t} = (x'_{1,t}, x'_{22,t})'$. The sample is divided into in-sample and out-of-sample portions. The in-sample observations span 1 to **R**. Letting **P** denote the number of 1-step ahead predictions, the out-of-sample observations span R + 1 through R + P. The total number of observations in the sample is R + P = T + 1.

Forecasts of y_{t+1}, t = R,…,T, are generated using two linear models of the form $x'_{i,t+1}\beta^*_i$, i = 1,2, each of which is estimated. Under the null, model 2 nests the restricted model 1 and hence model 2 includes k_2 excess parameters. Without loss of generality, let $\beta^*_2 = (\beta^{*'}_{1\ 1\times k_1},\ 0_{1\times k_2})'$. Under the alternative hypothesis, the k_2 restrictions are not true, and model 2 is correct.

The forecasts are *recursive*, 1-step ahead predictions. Under the recursive scheme, each model's parameters, β^*_i, i = 1,2, are estimated with added data as forecasting moves forward

Berkowitz and Giorgianni (1999), and Kilian (1999).

through time: for t = R,…,T, model i's prediction of y_{t+1}, $x'_{i,t+1}\hat{\beta}_{i,t}$, is created using the parameter estimate $\hat{\beta}_{i,t}$ based on data from 1 to t. The largest number of observations used to estimate the model parameters is then T = R + P − 1. Asymptotic results for forecasts based on *rolling* and *fixed* schemes are provided in Clark and McCracken (2000).[2]

We focus on 1-step ahead forecasts because, for multi-step forecasts, the asymptotic distributions of the tests generally appear to depend on the parameters of the data-generating process.[3] For practical purposes, such dependence eliminates the possibility of using asymptotically pivotal approximations to test for equal accuracy or encompassing. Given that most forecast comparisons include 1-step ahead results, our asymptotic results should be useful in many settings. For those researchers interested in multi-step horizons, bootstrap procedures, such as those developed in Ashley (1998) and Kilian (1999), may yield accurate inferences.

We denote the 1-step ahead forecast errors as $\hat{u}_{1,t+1} = y_{t+1} - x'_{1,t+1}\hat{\beta}_{1,t}$ and $\hat{u}_{2,t+1} = y_{t+1} - x'_{2,t+1}\hat{\beta}_{2,t}$ for models 1 and 2, respectively. The forecast encompassing tests are formed using these two sequences of P forecast errors. In all cases the out-of-sample statistics rely on sums of functions of these forecast errors. To simplify notation, for any variable z_{t+1} we let $\sum_t z_{t+1}$ denote the summation $\sum_{t=R}^{T} z_{t+1}$.

Before moving to the assumptions some final notation is needed. For i = 1,2 let $h_{i,t+1}(\beta_i) = (y_{t+1} - x'_{i,t+1}\beta_i)x_{i,t+1}$, $h_{i,t+1} = h_{i,t+1}(\beta_i^*)$, $q_{i,t+1} = x_{i,t+1}x'_{i,t+1}$ and $B_i = (Eq_{i,t+1})^{-1}$. For any (m×n) matrix A with elements $a_{i,j}$ and column vectors a_j, let vec(A) denote the (mn×1) vector

[2] West and McCracken (1998) discuss the rolling and fixed forecasting schemes. Clark and McCracken (2000) shows that, under the rolling scheme, the asymptotic distributions of the ENC-T and ENC-REG statistics considered below are non-standard for $\pi > 0$ and standard normal for $\pi = 0$. Under the fixed scheme, the ENC-T and ENC-REG tests are asymptotically standard normal for both $\pi > 0$ and $\pi = 0$.

$[a_1', a_2', ..., a_n']'$. Let W(s) denote a ($k_2 \times 1$) vector standard Brownian motion. Finally, under the null, $u_{1,t} = u_{2,t} \equiv u_t$.

Given the definitions and forecasting schemes described above, the following assumptions are used to derive the limiting distributions in Theorems 3.1-3.3. The assumptions are also sufficient for the results of McCracken (1999) when MSE is the measure of predictive ability. The assumptions are intended to be only sufficient, not necessary and sufficient.

<u>Assumption 1:</u> The parameter estimates $\hat{\beta}_{i,t}$, i = 1,2, t = R,…,T, satisfy $\hat{\beta}_{i,t} - \beta_i^* = B_i(t)H_i(t)$ where $B_i(t)H_i(t)$ equals $(t^{-1}\sum_{j=1}^{t} q_{i,j})^{-1}(t^{-1}\sum_{j=1}^{t} h_{i,j})$.

Our first assumption is that the parameters must be estimated by OLS. This restriction is imposed to ensure that the statistics in Theorems 3.1-3.3 are asymptotically pivotal. As in McCracken (1999), achieving a limiting distribution that does not depend upon the data-generating process requires that the loss function used to estimate the parameters be the same as the loss function used to measure predictive ability. Each of the statistics in Theorems 3.1-3.3 are functions of squared forecast errors. To achieve an asymptotically pivotal statistic the parameters must then be estimated using a squared error loss function.

<u>Assumption 2:</u> Let $U_t = [u_t, x_{2,t}' \text{-} Ex_{2,t}', h_{2,t}', vec(h_{2,t}h_{2,t}' \text{-} Eh_{2,t}h_{2,t}')', vec(q_{2,t} \text{-} Eq_{2,t})']'$. (a) $EU_t = 0$, (b) $Eq_{2,t} < \infty$ is p.d., (c) For some r > 4 U_t is uniformly L^r bounded, (d) For all t, $Eu_t^2 = \sigma^2 < \infty$, (e) For some r > d > 2, U_t is strong mixing with coefficients of size $-rd/(r - d)$, (f) Letting $\tilde{U}_t$ denote the vector of nonredundant elements of U_t, $\lim_{T \to \infty} T^{-1}E(\sum_{j=1}^{T}\tilde{U}_j)(\sum_{j=1}^{T}\tilde{U}_j)' = \Omega < \infty$ is p.d..

[3] Lutkepohl and Burda (1997) note similar difficulties associated with in-sample causality tests involving multi-step

Assumption 3: (a) $Eh_{2,t}h_{2,t}^{'} = \sigma^2 Eq_{2,t}$, (b) $E(h_{2,t} \mid h_{2,t-j}, q_{2,t-j}, j = 1,2,...) = 0$.

Assumptions 2 and 3 allow the application of an invariance principle and are sufficient for joint weak convergence of partial sums and averages of these partial sums to Brownian motion and integrals of Brownian motion. Assumption 2 is directly comparable to the assumptions in Hansen (1992) and hence we are able to apply his Theorems (2.1) and (3.1). Assumption 3 is also used to ensure that the limiting distribution does not depend upon the underlying data-generating process.

Assumption 4: $\lim_{P,R\to\infty} P/R = \pi$, $0 < \pi < \infty$, $\lambda \equiv (1+\pi)^{-1}$.

Assumption 4′: $\lim_{P,R\to\infty} P/R = \pi = 0$, $\lambda = 1$.

Assumptions 4 and 4′ introduce the alternative means by which the asymptotics are achieved. As in Ghysels and Hall (1990), West (1996), and White (2000) the limiting distribution results are derived by imposing a slightly stronger condition than simply that the sample size, T+1, becomes arbitrarily large. Here we impose the additional condition that *either* the numbers of in-sample (R) and out-of-sample (P) observations become arbitrarily large at the same rate (i.e. $P/R \to \pi > 0$) or the number of in-sample observations become arbitrarily large relative to the number of out-of-sample observations (i.e. $P/R \to 0$). As shown below, the asymptotics depend critically upon the value of π – that is, whether Assumption 4 or 4′ is made.

3. Tests and Asymptotic Distributions

We consider two standard forecast encompassing tests – those proposed by Ericsson

horizons.

(1992) and Harvey, Leybourne, and Newbold (1998) – as well as one new test. West and McCracken (1998) show that another standard test, proposed by Chong and Hendry (1986), can be asymptotically normal when applied to either nested or non-nested forecasts. In our Monte Carlo simulations, however, the power of the Chong and Hendry test was dominated by that of the tests described below. A related test, the out-of-sample causality statistic developed by Chao, Corradi, and Swanson (2000), is also asymptotically normal.

3.1 The ENC-T Test

Drawing on the methodology of Diebold and Mariano (1995), Harvey, Leybourne, and Newbold (1998) propose a test of encompassing that uses a t-statistic for the covariance between $\hat{u}_{1,t+1}$ and $\hat{u}_{1,t+1} - \hat{u}_{2,t+1}$. Let $c_{t+1} = \hat{u}_{1,t+1}(\hat{u}_{1,t+1} - \hat{u}_{2,t+1}) = \hat{u}^2_{1,t+1} - \hat{u}_{1,t+1}\hat{u}_{2,t+1}$ and $\bar{c} = P^{-1}\sum_t c_t$. Their encompassing test, denoted ENC-T, is formed as

$$\text{ENC-T} = (P-1)^{1/2}\frac{\bar{c}}{\sqrt{P^{-1}\sum_t (c_{t+1}-\bar{c})^2}} = (P-1)^{1/2}\frac{P^{-1}\sum_t(\hat{u}^2_{1,t+1} - \hat{u}_{1,t+1}\hat{u}_{2,t+1})}{\sqrt{P^{-1}\sum_t(\hat{u}^2_{1,t+1} - \hat{u}_{1,t+1}\hat{u}_{2,t+1})^2 - \bar{c}^2}}. \quad (1)$$

The term in front is $(P-1)^{1/2}$ rather than $P^{1/2}$ because we calculate the test using standard regression methods (we regress c_{t+1} on a constant). Under the null that model 1 forecast encompasses model 2, the covariance between $u_{1,t}$ and $u_{1,t} - u_{2,t}$ will be less than or equal to 0. Under the alternative that model 2 contains added information, the covariance should be positive. Hence the ENC-T test, and the other encompassing tests described below, are one-sided.

Theorem 3.1: (a) Let Assumptions 1-4 hold. For ENC-T defined in (1), ENC-T $\rightarrow_d$ $\Gamma_1/(\Gamma_2)^{1/2}$ where $\Gamma_1 = \int_\lambda^1 s^{-1}W'(s)dW(s)$ and $\Gamma_2 = \int_\lambda^1 s^{-2}W'(s)W(s)ds$. (b) Let Assumptions 1-3 and 4′ hold. ENC-T $\rightarrow_d$ N(0, 1).

While West (1996) shows that the ENC-T statistic can be asymptotically normal for any value of $\pi \geq 0$ when applied to non-nested forecasts, this is not the case when the models are nested. In Theorem 3.1 (a), we show that if $\pi > 0$, the ENC-T statistic has a nonstandard limiting distribution. Although this null limiting distribution does not depend upon the data-generating process, it does depend on two parameters. The first is the number of excess parameters k_2, which arises because the vector Brownian motion, W(s), is $(k_2 \times 1)$. The second parameter is π, which affects the range of integration on each of the stochastic integrals through λ. In Theorem 3.1 (b), however, we show that if $\pi = 0$, the ENC-T statistic is limiting standard normal.[4]

We provide a selected set of numerically-generated asymptotic critical values for the ENC-T statistic, when $\pi > 0$, in the upper panel of Table 1.[5] The reported critical values are percentiles of 5000 independent draws from the distribution of $\Gamma_1/(\Gamma_2)^{1/2}$ for a given value of k_2 and π. In generating these draws, the necessary k_2 Brownian motions are simulated as random walks each using an independent sequence of 10,000 i.i.d. $N(0,T^{-1/2})$ increments, and the integrals are emulated by summing the relevant weighted quadratics of the random walks.

The asymptotic critical values for $\pi > 0$ clearly differ from the standard normal critical values that are appropriate when $\pi = 0$. For example, with $\pi = 1$ and $k_2 = 1$, the 90th percentile of the asymptotic distribution is 0.955, compared to 1.282 for the standard normal distribution. As π declines, the asymptotic critical values rise gradually, but remain somewhat different from standard normal values. With $\pi = 0.2$ and $k_2 = 1$, for instance, the 90th percentile of the asymptotic distribution is 1.002.

[4] Rather than following from the standard application of a cental limit theorem, this limiting normality result is analogous to the (finite-sample) unconditional normality of t-statistics from linear regressions with stochastic regressors and i.i.d. normal disturbances. Clark and McCracken (2000) provides some additional detail.

3.2 The ENC-REG Test

The forecast encompassing test proposed by Ericsson (1992) is a regression-based variant of the ENC-T test. The test statistic, denoted ENC-REG, is the t-statistic associated with the coefficient α_1 from the OLS regression $\hat{u}_{1,t+1} = \alpha_1(\hat{u}_{1,t+1} - \hat{u}_{2,t+1})$ + error term, which can be expressed as

$$\text{ENC-REG} = (P-1)^{1/2}\frac{P^{-1}\sum_t \hat{u}_{1,t+1}(\hat{u}_{1,t+1} - \hat{u}_{2,t+1})}{\sqrt{P^{-1}\sum_t(\hat{u}_{1,t+1} - \hat{u}_{2,t+1})^2(P^{-1}\sum_t \hat{u}^2_{1,t+1}) - \bar{c}^2}}. \qquad (2)$$

Theorem 3.2: Let Assumptions 1-3 and either 4 or 4′ hold. For ENC-REG defined in (2) and ENC-T defined in (1), ENC-REG – ENC-T = $o_p(1)$.

While the ENC-REG statistic, like the ENC-T statistic, can be asymptotically normal for any value of $\pi \geq 0$ when applied to non-nested forecasts, this is not the case when the models are nested. Theorem 3.2 states that, with nested models, regardless of whether $\pi > 0$ or $\pi = 0$, ENC-REG and ENC-T are asymptotically equivalent under the null.[6] Therefore, the asymptotic distribution of ENC-REG is non-standard when $\pi > 0$ and standard normal when $\pi = 0$.

3.3 A New Encompassing Test

Because the population prediction errors from models 1 and 2 are exactly the same under the null (making c_{t+1}, in population, identically 0) the sample variances in the denominators of the ENC-T and ENC-REG statistics (1) and (2) are, heuristically, 0. This feature of the ENC-T and ENC-REG statistics may adversely affect the small-sample properties of the tests. Therefore, we propose a variant of the ENC-T and ENC-REG statistics in which $\bar{c}$ (the

[5] Clark and McCracken (2000) provides more detailed tables, covering additional values of k_2 and π as well as the rolling and fixed forecasting schemes.

covariance between $\hat{u}_{1,t+1}$ and $\hat{u}_{1,t+1} - \hat{u}_{2,t+2}$) is scaled by the variance of one of the forecast errors rather than an estimate of the variance of $\bar{c}$.[7] This statistic, which we refer to as the ENC-NEW test, takes the form

$$\text{ENC-NEW} = P \cdot \frac{\bar{c}}{\text{MSE}_2} = P \cdot \frac{P^{-1}\sum_t (\hat{u}_{1,t+1}^2 - \hat{u}_{1,t+1}\hat{u}_{2,t+1})}{P^{-1}\sum_t \hat{u}_{2,t+1}^2}. \qquad (3)$$

Theorem 3.3: (a) Let Assumptions 1-4 hold. For ENC-NEW defined in (3) and Γ_1 defined in Theorem 3.1, ENC-NEW $\to_d \Gamma_1$. (b) Let Assumptions 1-3 and 4′ hold. ENC-NEW $\to_p 0$.

As with the ENC-T and ENC-REG statistics, if $\pi > 0$ the limiting distribution of ENC-NEW is non-normal when the forecasts are nested under the null. The limiting distribution of ENC-NEW is also asymptotically pivotal and dependent on the parameters k_2 and π. We provide a selected set of asymptotic critical values for the ENC-NEW statistic in the lower panel of Table 1. These values were generated numerically using the limiting distribution in Theorem 3.3 (a).

Theorem 3.3 (b) shows that, if $\pi = 0$, the limiting distribution of the ENC-NEW statistic is degenerate – not standard normal as in the case of the ENC-T and ENC-REG tests. As noted by Chong and Hendry (1986), when $\pi = 0$ the parameter estimates are essentially ‘known’ before the out-of-sample period begins. We would then expect the numerator of the ENC-NEW statistic to behave like its population counterpart, which is 0. The same logic does not apply to the ENC-T and ENC-REG statistics because both the numerator and the denominator of these

[6] Similarly, McCracken (1999) shows that the MSE-REG and MSE-T tests included in our Monte Carlo simulations are asymptotically equivalent.

[7] We use MSE_2 in the denominator of the ENC-NEW statistic for consistency with the formulation of McCracken’s MSE-F test. Replacing MSE_2 with MSE_1 leaves the asymptotic null distribution the same and produces finite-sample size and power results very similar to those reported in section 4.

statistics are converging to zero at the same rate.[8]

4. Monte Carlo Results

The small-sample properties of the encompassing tests described in Section 3, as well as some tests of equal MSE, are evaluated using simulations of bivariate data-generating processes. The equal MSE tests are those for which McCracken (1999) derives the asymptotic distributions: an F-type test proposed by McCracken (MSE-F), a t-test proposed by Diebold and Mariano (1995) (MSE-T), and the Granger and Newbold (1977) t-test (MSE-REG).[9] While the analysis is focused on testing ex-ante forecasts for equal accuracy and encompassing, for the sake of comparison we also provide results for the standard full-sample F-test of Granger causality (GC).

In these simulations, we compare the predictive ability of an AR model (model 1) with that from a VAR model (model 2). The presented results are based on data generated using standard normal disturbances. The results are essentially unchanged when the disturbances are drawn from the heavier-tailed t(6) distribution considered by Diebold and Mariano (1995), Harvey, Leybourne, and Newbold (1997, 1998), and Clark (1999). The forecasts in our presented results are recursive; using rolling and fixed forecasts generally produces the same results.[10]

4.1 Experiment Design

In the presented results, data are generated using two different models. The first, denoted DGP-I, takes the form

[8] Clark and McCracken (2000) shows that, if $\pi = 0$, the numerator and denominator terms of the ENC-T and ENC-REG statistics are each $o_p((P/R)^{1/2})$.

[9] Because the models are nested, the null hypothesis is $Eu^2_{1,t+1} \leq Eu^2_{2,t+1}$ and the alternative is $Eu^2_{1,t+1} > Eu^2_{2,t+1}$. The alternative is one-sided because, if the restrictions imposed on model 1 are not true, there is no reason to expect forecasts from model 1 to be superior to those from model 2.

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} 0.3 & b \\ 0 & 0.5 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} u_{y,t} \\ u_{x,t} \end{pmatrix}. \tag{4}$$

The second, denoted DGP-II, takes the form

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} 0.3 & b \\ 0.7 & -0.5 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} 0.3 & 0 \\ 0.3 & 0 \end{pmatrix} \begin{pmatrix} y_{t-2} \\ x_{t-2} \end{pmatrix} + \begin{pmatrix} u_{y,t} \\ u_{x,t} \end{pmatrix}. \tag{5}$$

In both cases, y_t is the predictand, x_t is an auxiliary variable, and the disturbances are i.i.d. standard normal random variates. To evaluate size, the coefficient b is set at 0. In this case, the AR and VAR models have equal MSE and forecasts from the AR model encompass those from the VAR. To evaluate power, b is set at 0.1 and 0.2. In these power experiments, the VAR forecasts of y_{t+1} have lower MSE than the AR forecasts, and the AR forecast does not encompass the VAR forecast. Simulations based on several other DGPs, including the empirical inflation and unemployment model considered in Section 5, produced similar results.

In each Monte Carlo simulation we generate R + P + 4 observations. The additional four observations allow for data-determined lag lengths in the forecasting models. After drawing initial observations from the unconditional normal distribution implied by the DGP, the remaining observations are constructed iteratively using the autoregressive model structure and draws of the error terms from the standard normal distribution. After reserving observations 1 through 4 to allow for a maximum of four data-determined lags, the in-sample period spans observations 5 through R + 4. The estimated forecasting models are used to form P 1-step ahead, recursive predictions, spanning observations R + 5 through R + P + 4.

We have generated results based on a variety of methods for determining the lag lengths of the estimated models. Specifically, we consider simply fixing the lag length at the true order

[10] The key exception is that, with fixed forecasts, comparing the ENC-T and ENC-REG tests against standard normal critical values does not produce undersized tests because, with fixed forecasts, the tests are standard normal

of the DGP as well as using AIC, SIC, and a last significant lag criterion to determine the optimal lag.[11] In employing the data-based methods, we use a particular criterion to determine the optimal lag length for the estimated VAR model and then impose the same order on the estimated AR model.[12] The lag lengths of the models used for forecasting were determined using only the in-sample portion of the data. However, the estimated model underlying each GC test uses a lag length determined from the full sample of R + P observations.

In most, although not all, instances, our basic results are not sensitive to the lag selection method. Accordingly, we focus the discussion on results based on setting the lag lengths at the true order. We then include a brief discussion comparing results across lag selection methods. Because the AIC and last significant lag criteria yield similar results, this brief discussion is focused on the performance of AIC and SIC.

In our Monte Carlo experiments, the ENC-NEW and MSE-F test results are based on comparing the statistics against the asymptotic critical values provided in Table 1 and McCracken (1999), respectively. For the ENC-T, ENC-REG, MSE-T and MSE-REG tests, we report two sets of results: one based on the asymptotic critical values reported in Table 1 or McCracken (1999) and another based on standard normal critical values. For these four tests, the true asymptotic critical values are standard normal only when $\pi = 0$. In our experiments, $\hat{\pi} \equiv$ P/R is non-zero, but sometimes small. Our experiments address whether using a standard normal approximation is accurate when $\hat{\pi}$ is small.

Results are reported for empirically relevant combinations of P and R such that $\hat{\pi}$ takes the values 0.1, 0.2, 0.4, 1.0, 2.0, 3.0, or 5.0. Specifically, we use R = 50 with P = 100, 150, and

for all π.

[11] We have also examined results based on simply fixing the lag length at 4, which yields similar results, except that all tests have lower power. Our last significant lag criterion is the general-to-specific Wald test described in Hall (1994) and Ng and Perron (1995).

200; R = 100 with P = 10, 20, 40, 100, and 200; and R = 200 with P = 20, 40, 80, and 200.

4.2 Size Results

Table 2 presents the empirical sizes of Granger causality, equal forecast accuracy, and forecast encompassing tests for data from DGP-I and DGP-II, using a nominal size of 10%.[13] Table 3 presents a selected set of comparable results based on data-determined lag lengths. Three general results are evident from these tables.

Size result 1. In most settings the post-sample tests have reasonable finite-sample size properties when compared against asymptotic critical values for $\pi = \hat{\pi} \equiv P/R$. Specifically, the MSE-F, MSE-REG, ENC-NEW, and ENC-REG tests perform well, suffering only slight size distortions in finite samples. For example, Table 2 shows that, with DGP-I, R = 100, and P = 20, these four tests have empirical sizes of 10.7%, 11.7%, 11.0%, and 11.8%, respectively. While the MSE-T and ENC-T statistics also perform reasonably well, when P is small the tests suffer slightly greater distortions than do the MSE-REG and ENC-REG tests. For instance, using DGP-I, R = 100, and P = 10, the MSE-T test has an actual size of 15.4% while MSE-REG has an actual size of 13.0%. The better performance of MSE-REG and ENC-REG likely stems from the regression forms of the tests using more precise variance estimates.[14] For example, the variance term in the denominator of the ENC-REG test (2) uses a product of second moments, whereas the ENC-T test (1) uses a sample fourth moment.

In general, given R, the size distortions of the post-sample tests fall as P rises. For instance, when data are generated using DGP-I with R = 100 and P = 10, Table 2 shows that actual size ranges from 11.3% to 15.4%. When P increases to 100, actual size ranges from

[12] Setting the lag length based on just the equation for y_t yields similar results.

[13] The results are generally the same at a nominal size of 5%.

[14] We also find that MSE-REG and ENC-REG have better size properties than MSE-T and ENC-T in simulations with t(6)-distributed innovations.

10.4% to 11.0%. Note that the absence of any size distortions in the results for R = 50 reflects the fact that P is large.

Size result 2. Comparing the MSE-T, MSE-REG, ENC-T, and ENC-REG tests against standard normal critical values generally leads to too-infrequent rejections. The problem is most severe for the MSE-T and MSE-REG tests. For instance, using DGP-II, R = 100, P = 20 and standard normal critical values, Table 2 shows the MSE-T and MSE-REG tests yield sizes of 5.6% and 4.7%, respectively. In accordance with the theory, for a given R, the tests become more undersized as P rises. In the same example, but with P = 100, the sizes of the MSE-T and MSE-REG tests fall to 1.4% and 1.3%, respectively.

For small values of $\hat{\pi} \equiv P/R$, whether the asymptotic critical values for $\pi = \hat{\pi}$ or standard normal critical values associated with $\pi = 0$ provide better finite-sample results is largely a matter of individual judgment.[15] As the results in Table 2 show, when P is relatively small, using the critical values in Table 1 and McCracken (1999) yields slightly over-sized tests, while using standard normal critical values yields slightly under-sized tests. Again, though, standard normal critical values are a better approximation for the empirical distributions of the ENC-T and ENC-REG statistics than of the MSE-T and MSE-REG statistics.

This second size result, as well as the first, continues to hold when data-based methods are used to determine the lag length, as long as lag selection is reasonably accurate. For example, with DGP-I and R = 100, the AIC and SIC select the true lag about 86% and 99.8% of the time, respectively. Accordingly, as evident from a comparison between the data-determined lag length results in Table 3 and the fixed-lag results in Table 2, with DGP-I there are few differences in the sizes of the forecast tests across lag selection methods. Similarly, with DGP-II and R = 200, AIC is sufficiently accurate that the sizes of the forecast tests are essentially the same when the lag is data-determined as when it is fixed at the true order. Our final size result addresses some differences that arise when lag selection is less accurate.

Size result 3. When data-based lag selection is sufficiently imprecise, size performance deteriorates.

In the case of DGP-II, the true model for y_t is an AR(2). However, because the population correlation between x_{t-1} and y_{t-2} is large (0.57), data-based procedures often select a lag of 1 (for an estimated VAR in x_t and y_t). For example, when R = 100, the AIC selects a lag of 1 in about 13% of the DGP-II simulations, while the SIC selects a lag of 1 with a frequency of 67%. When R = 200, the AIC selects a lag of 1 with a frequency of just 0.6%, while the SIC selects a lag of 1 with a frequency of 24%.

The difficulties in selecting the lag length in DGP-II simulations create modest-to-substantial size distortions in the forecast tests, with the SIC producing the largest distortions.[16] Table 3 shows that, when R = 100 and P = 40, using the AIC to determine lag length makes the size of the ENC-NEW test 16.2%, while using the SIC makes the size 34.4%. While increasing R to 200 eliminates the distortions in AIC-based tests (compared to tests based on the true lag length), modest distortions remain in SIC-based tests. For instance, the ENC-NEW test has a size of 22.5% when R = 200, P = 40, and the lag is selected using the SIC.

Our analysis of different lag selection procedures also shows that data-based methods can create size distortions in the GC test that rival and sometimes exceed those of the post-sample tests. For example, as reported in Table 3, in the experiment using DGP-I, R = 100, P = 40, and

[15] Some unreported results show that while the 90th and 95th percentiles of the empirical distribution very roughly approximate the corresponding percentiles of the standard normal distribution, the null of normality of the empirical distribution of each tests is strongly rejected for P/R = 0.1 or 0.2.

AIC, the GC test has empirical size of 13.3%, compared to a range of 10.2% to 12.0% for the post-sample tests.[17] Similarly, in the experiment with DGP-II, R = 200, and P = 40, the AIC-based GC test has size of 13.0%, compared to a range of 10.6% to 12.0% for the post-sample tests. While using the SIC to select the lag length makes the GC test correctly sized in experiments with DGP-I, in experiments with DGP-II the SIC-based GC test often suffers size distortions that rival or exceed those of the post-sample tests. For example, with R = 100 and P = 40, the GC test has size of 33.1%, compared to a range of 25.2% to 34.4% for the post-sample tests.

Some other evidence suggests that the size advantage of post-sample tests may be even larger when more data mining is involved in choosing the lag length of the VAR in x_t and y_t. Yet another, more data-intensive approach to model selection is to allow the lags on y_t and x_t in the nesting equation for y_t (i.e., model 2) to differ, and then choose the lag combination that minimizes the AIC for that equation. Using this approach to lag selection, when the DGP is DGP-I, R = 100, and P = 20, the GC test has actual size of 20.3%, while the MSE-F and ENC-NEW tests have size of 11.6% and 13.5%, respectively.

4.3 Power Results

Tables 4 and 5 present results on the power of forecast encompassing, equal forecast accuracy, and Granger causality tests. Because the tests are, to varying degrees, subject to size distortions, the reported power figures are based on empirical critical values and therefore size-adjusted.[18] The actual size of the tests is 10%. Two general results are evident in Tables 4 and

[16] The distortions do not decline as P rises. For instance, when R = 100 and P = 100, the AIC- and SIC-based versions of the ENC-NEW test have sizes of 18.1% and 47.0%, respectively.

[17] For both R = 100 and R = 200 the size of the AIC-based GC test remains at about 13% as P is increased.

[18] In results allowing data-determined lags in a given experiment, the test statistic in simulation i, for which the selected lag is j, is compared against the distribution of test statistics from the set of corresponding size simulations in which the lag was selected to be j.

5.

Power result 1. The small-sample powers of the tests generally permit some simple rankings: ENC-NEW > MSE-F, ENC-T, ENC-REG > MSE-T, MSE-REG. In our experiments, the ENC-NEW test is clearly the most powerful out-of-sample test of predictive ability. In some settings, the power of the ENC-NEW statistic rivals the power of the GC test, even though the GC test is based on many more observations (R + P rather than P). For example, as shown in the lower panel of Table 4, in simulations with DGP-II, b = 0.1, R = 100, and P = 40, the ENC-NEW test has power of 26.4%, comparable to the GC test's power of 31.0%. The MSE-F, ENC-T, and ENC-REG tests are less powerful than the ENC-NEW test. Using the experiment of the previous example, the MSE-F, ENC-T, and ENC-REG tests have power of 22.8%, 22.3%, and 22.8%, respectively. The MSE-T and MSE-REG tests are less powerful than these tests.

Power result 2. Increasing the number of observations affects the powers of the tests in two basic ways. First, holding P fixed, the powers of the post-sample tests tend to rise with R, although more for some tests than others.[19] For instance, as shown in the upper panel of Table 4, with DGP-I and P = 40, the power of the ENC-NEW test rises from 33.2% when R = 100 to 41.4% when R = 200. Second, when R is held fixed, power rises with P. For example, Table 5 shows that, in experiments with DGP-I, R = 100 and b = 0.2, the power of the MSE-F test rises from 41.1% when P = 10 to 77.7% when P = 100. The powers of the three tests for equal MSE converge as P becomes large, and the same happens for the three encompassing tests.

Our two key power results still hold when data-based methods are used to determine the lag length. With DGP-I, SIC-based power is virtually the same as when the lag is set at the true order of the DGP; AIC-based power is the same to slightly lower. For example, in the DGP-I experiment with R = 100, P = 40, and b = 0.1, using the SIC yields power of 33.2% for the ENC-

NEW test, while using the AIC yields power of 31.2%. With DGP-II, our two key power results hold for all lag selection methods, but the lag selection problems discussed above give the SIC a power advantage. For instance, in the DGP-II experiment with R = 100, P = 40, and b = 0.1, the SIC-based ENC-NEW test has power of 42.0% for the ENC-NEW test, while the AIC-based test has power of 27.8%.

5. Empirical Example

In this section we use tests of forecast encompassing, equal forecast accuracy, and Granger causality to determine whether the prime-age male unemployment rate is useful in predicting core CPI inflation. Cecchetti (1995), Staiger, Stock, and Watson (1997), and Stock and Watson (1999) are recent examples of studies in the long literature on this basic question.

Our quarterly data, which begin in 1957:Q1, are divided into in-sample and out-of-sample portions so as to produce a $\hat{\pi} \equiv P/R$ value for which this paper and McCracken (1999) report corresponding asymptotic critical values. After we allow for data differencing and a maximum of four data-determined lags, the in-sample period spans 1958:Q3-1987:Q1, for a total of 115 observations. The out-of-sample period spans 1987:Q2-1998:Q3, yielding a total of P = 46 1-step ahead predictions. For this split, $\hat{\pi} = 0.4$.

Consistent with the results of augmented Dickey-Fuller tests for unit roots, our model variables are the change in inflation and the change in the unemployment rate. Over the in-sample period, AIC is minimized at two lags for both the AR and the VAR. The test statistics are compared against asymptotic critical values for $\pi = 0.4$ from Table 1 and McCracken (1999) and empirical critical values generated from Monte Carlo simulations of the estimated inflation-unemployment model in which the null of no causality from unemployment to inflation is

[19] However, in a few cases, the powers of the MSE-T and MSE-REG tests decline as R rises given P.

imposed. As can be seen from the critical values reported in the lower panel of Table 6, the asymptotic critical values for $\pi = 0.4$ provide a good approximation to the empirical critical values – a better approximation than provided by the standard normal critical values that are appropriate for $\pi = 0$.

The upper panel of Table 6 reports in-sample estimates of an AR(2) fit to changes in core CPI inflation and a VAR(2) fit to changes in core CPI inflation and prime-age male unemployment. In the in-sample model estimates, unemployment clearly has predictive power for inflation. Moreover, the full-sample GC test reported in the lower panel of the table strongly rejects the null of no causality from unemployment to inflation.

Although weaker, the out-of-sample evidence also indicates unemployment has predictive power for inflation. As reported in the lower panel of Table 6, all of the encompassing tests indicate that the change in unemployment has predictive content for the change in inflation. The ENC-NEW test strongly rejects the null that the AR forecast encompasses the VAR forecast. The ENC-REG test clearly rejects, while the ENC-T test marginally rejects. None of the tests for equal MSE reject the null of equal accuracy.

Two factors may account for the difference in the strength of the in-sample and post-sample evidence. One is simply power differences – some of the post-sample tests may not be powerful enough to pick up unemployment’s predictive content. The Monte Carlo results in Section 4 indicate that the power of equal forecast accuracy tests, such as MSE-F, lag behind the power of encompassing counterparts like the ENC-NEW test, which has power rivaling that of the GC test. The second factor is model instabilities. Neither the AR model for inflation nor the VAR pass the supremum Wald or exponential Wald tests for stability developed in Andrews (1993) and Andrews and Ploberger (1994), respectively.

6. Conclusions

In this paper we first derive the limiting distributions of two standard tests and one new test of forecast encompassing applied to 1-step ahead predictions from nested linear models. We show that the tests have non-standard distributions when the number of observations used to generate initial estimates of the models and the number of forecast observations increase at the same rate. We then provide numerically-generated critical values for these distributions. We also show that the two standard tests are limiting standard normal when the number of forecasts increases at a slower rate than the number of observations used in the initial model estimates.

We then use Monte Carlo experiments to examine the finite-sample size and size-adjusted power of equal accuracy and encompassing tests. These experiments yield three essential results. First, the post-sample tests are, in general, reasonably well sized when the critical values provided in this paper are used. Second, when standard normal critical values are used the post-sample tests are undersized. Third, the encompassing test proposed in this paper (the ENC-NEW statistic defined in equation (3)) is most powerful.

In the final part of our analysis, we find that the post-sample tests provide mixed, but suggestive, evidence on the predictive content of unemployment for inflation in the U.S. Although all of the equal forecast accuracy tests fail to reject the null that unemployment has no predictive content for inflation, each of the encompassing tests indicates that unemployment does have predictive power. Since encompassing tests appear to have a power advantage in finite samples, unemployment probably does have some predictive value.

References

Amano, R.A. and S. van Norden, 1995, Terms of trade and real exchange rates: The Canadian evidence, Journal of International Money and Finance 14, 83-104.

Andrews, D.W.K., 1993, Tests for parameter instability and structural change with unknown change point, Econometrica 61, 821-56.

Andrews, D.W.K. and W. Ploberger, 1994, Optimal tests when a nuisance parameter is present only under the alternative, Econometrica 62, 1383-1414.

Ashley, R., 1998, A new technique for postsample model selection and validation, Journal of Economic Dynamics and Control 22, 647-665.

Ashley, R., C.W.J. Granger and R. Schmalensee, 1980, Advertising and aggregate consumption: An analysis of causality, Econometrica 48, 1149-67.

Berkowitz, J. and L. Giorgianni, 1999, Long-horizon exchange rate predictability, Review of Economics and Statistics, Forthcoming.

Bram, J. and S. Ludvigson, 1998, Does consumer confidence forecast household expenditure? A sentiment horse race, *Economic Policy Review*, Federal Reserve Bank of New York, June, 59-78.

Cecchetti, S.G., 1995, Inflation indicators and inflation policy, NBER macroeconomics annual, 189-219.

Chao, J., V. Corradi and N. Swanson, 2001, An out of sample test for Granger causality, Macroeconomic Dynamics, forthcoming.

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.

Chong, Y.Y. and D.F. Hendry, 1986, Econometric evaluation of linear macroeconomic models, Review of Economic Studies 53, 671-90.

Clark, T.E., 1999, Finite-sample properties of tests for equal forecast accuracy, Journal of Forecasting 18, 489-504.

Clark, T.E. and M.W. McCracken, 2000, Not-for-publication appendix to ‘Tests of equal forecast accuracy and encompassing for nested models’, manuscript, Federal Reserve Bank of Kansas City (available from www.kc.frb.org).

Corradi, V., N.R. Swanson and C. Olivetti, 1999, Predictive ability with cointegrated variables, manuscript, Texas A & M University.

Diebold, F.X. and R.S. Mariano, 1995, Comparing predictive accuracy, Journal of Business and Economic Statistics 13, 253-63.

Diebold, F.X. and G.D. Rudebusch, 1991, Forecasting output with the composite leading index: A real time analysis, Journal of the American Statistical Association 86, 603-610.

Ericsson, N.R., 1992, Parameter constancy, mean square forecast errors, and measuring forecast performance: An exposition, extensions, and illustration, Journal of Policy Modeling 14, 465-95.

Ghysels, E. and A. Hall, 1990, A test for structural stability of Euler conditions parameters estimated via the generalized method of moments estimator, International Economic Review 31, 355-64.

Granger, C.W.J. and P. Newbold, 1977, Forecasting Economic Time Series (Academic Press, Orlando, FL).

Hall, A., 1994, Testing for a unit root in time series with pretest databased model selection, Journal of Business and Economics Statistics 12, 461-70.

Hansen, B.E., 1992, Convergence to stochastic integrals for dependent heterogeneous processes, Econometric Theory 8, 489-500.

Harvey, D.I., S.J. Leybourne and P. Newbold, 1997, Testing the equality of prediction mean squared errors, International Journal of Forecasting 13, 281-91.

Harvey, D.I., S.J. Leybourne and P. Newbold, 1998, Tests for forecast encompassing, Journal of Business and Economic Statistics 16, 254-59.

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

Krueger, J.T. and K.N. Kuttner, 1996, The fed funds futures rate as a predictor of Federal Reserve policy, Journal of Futures Markets 16, 865-79.

Lutkepohl, H. and M.M. Burda, 1997, Modified Wald tests under nonregular conditions, Journal of Econometrics 78, 315-332.

Mark, N.C., 1995, Exchange rates and fundamentals: Evidence on long-horizon predictability, American Economic Review 85, 201-18.

McCracken, M.W., 2000, Robust out of sample inference, Journal of Econometrics 99, 195-223.

McCracken, M.W., 1999, Asymptotics for out-of-sample tests of causality, manuscript, Louisiana State University.

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

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

Ng, S. and P. Perron, 1995, Unit root rests in ARMA models with data-dependent methods for the selection of the truncation lag, Journal of the American Statistical Association 90, 268-81.

Staiger, D., J.H. Stock and M.W. Watson, 1997, The NAIRU, unemployment and monetary policy, Journal of Economic Perspectives 11, 33-49.

Stock, J.H. and M.W. Watson, 1999, Forecasting inflation, Journal of Monetary Economics 44, 293-335.

West, K.D., 1996, Asymptotic inference about predictive ability, Econometrica 64, 1067-84.

West, K.D., 2000a, Tests for forecast encompassing when forecasts depend on estimated regression parameters, Journal of Business and Economic Statistics, forthcoming.

West, K.D., 2000b, Encompassing tests when no model is encompassing, manuscript, University of Wisconsin.

West, K.D. and M.W. McCracken, 1998, Regression-based tests of predictive ability, International Economic Review 39, 817-40.

White, H., 2000, A reality check for data snooping, Econometrica 68, 1067-84.

Table 1

Percentiles of the ENC-T, ENC-REG, and ENC-NEW Statistics, $\pi > 0$: Recursive Scheme

ENC-T and ENC-REG

					$\pi =$			
k_2	%-ile	.1	.2	.4	1.0	2.0	3.0	5.0
1	0.95	1.422	1.360	1.338	1.331	1.322	1.329	1.336
	0.90	1.056	1.002	1.005	.955	.939	.937	.922
2	0.95	1.505	1.467	1.445	1.413	1.443	1.409	1.380
	0.90	1.166	1.101	1.086	1.066	1.035	1.034	1.028
3	0.95	1.574	1.525	1.529	1.476	1.473	1.469	1.436
	0.90	1.227	1.138	1.105	1.113	1.114	1.083	1.074
4	0.95	1.594	1.596	1.552	1.463	1.481	1.474	1.445
	0.90	1.219	1.175	1.192	1.132	1.111	1.091	1.090

ENC-NEW

					$\pi =$			
k_2	%-ile	.1	.2	.4	1.0	2.0	3.0	5.0
1	0.95	.520	.744	1.079	1.584	2.085	2.374	2.685
	0.90	.335	.473	.685	.984	1.280	1.442	1.609
2	0.95	.766	1.028	1.481	2.234	2.889	3.293	3.627
	0.90	.524	.716	1.019	1.471	1.914	2.074	2.428
3	0.95	.940	1.273	1.865	2.709	3.564	3.989	4.384
	0.90	.686	.890	1.285	1.905	2.366	2.664	3.132
4	0.95	1.060	1.526	2.181	3.007	3.894	4.542	4.957
	0.90	.776	1.062	1.528	2.169	2.727	3.032	3.513

Notes:
1. The test statistics ENC-T, ENC-REG, and ENC-NEW are defined in Section 3.
2. The upper panel of Table 1 reports estimates of the 90th and 95th percentiles of the asymptotic distribution of both the ENC-T and ENC-REG statistics when the recursive scheme is used and $\pi > 0$. The lower panel reports the corresponding percentiles of the asymptotic distribution of the ENC-NEW statistic.
3. The estimates were constructed based upon 5,000 simulated draws from the relevant distribution for given values of k_2 and π. See Section 3.1 of the text for further detail on how the simulations were conducted.

Table 2

Empirical Size

Recursive Forecasts

Nominal Size = 10%

	$R = 50$			$R = 100$					$R = 200$			
	P=100	P=150	P=250	P=10	P=20	P=40	P=100	P=200	P=20	P=40	P=80	P=200
DGP-I												
Tests Compared Against Asymptotic Critical Values for $\pi = \hat{\pi} \equiv P/R$												
MSE-F	.098	.097	.091	.113	.107	.100	.106	.099	.107	.104	.096	.102
MSE-T	.100	.097	.092	.154	.132	.118	.107	.099	.133	.121	.110	.103
MSE-REG	.097	.096	.092	.130	.117	.110	.104	.098	.119	.111	.104	.101
ENC-NEW	.104	.104	.105	.118	.110	.103	.105	.100	.111	.106	.099	.101
ENC-T	.109	.105	.101	.151	.135	.118	.110	.104	.131	.123	.108	.104
ENC-REG	.102	.102	.099	.128	.118	.107	.106	.101	.117	.113	.102	.101
GC	.103	.104	.101	.103	.102	.103	.104	.102	.103	.099	.101	.101
Tests Compared Against Standard Normal Critical Values												
MSE-T	.011	.007	.004	.088	.060	.040	.020	.010	.073	.052	.035	.018
MSE-REG	.010	.007	.004	.071	.049	.034	.018	.009	.062	.045	.032	.016
ENC-T	.061	.059	.054	.110	.090	.076	.065	.057	.093	.079	.068	.060
ENC-REG	.056	.055	.052	.093	.076	.068	.060	.055	.082	.071	.063	.057
DGP-II												
Tests Compared Against Asymptotic Critical Values for $\pi = \hat{\pi} \equiv P/R$												
MSE-F	.094	.094	.095	.114	.112	.106	.102	.097	.106	.108	.104	.101
MSE-T	.095	.094	.095	.139	.126	.118	.104	.099	.123	.116	.110	.102
MSE-REG	.095	.093	.095	.120	.114	.110	.102	.098	.109	.108	.107	.101
ENC-NEW	.102	.107	.104	.123	.115	.108	.105	.100	.110	.110	.102	.104
ENC-T	.105	.104	.102	.137	.130	.118	.106	.106	.121	.119	.110	.103
ENC-REG	.100	.100	.100	.119	.115	.107	.101	.103	.108	.109	.104	.100
GC	.097	.099	.100	.100	.098	.101	.100	.099	.099	.099	.098	.099
Tests Compared Against Standard Normal Critical Values												
MSE-T	.007	.004	.002	.082	.056	.033	.014	.006	.069	.049	.029	.012
MSE-REG	.005	.003	.002	.068	.047	.028	.013	.005	.061	.043	.026	.011
ENC-T	.069	.068	.065	.115	.098	.084	.074	.067	.101	.088	.077	.070
ENC-REG	.065	.064	.063	.100	.085	.075	.069	.065	.090	.079	.072	.067

Notes:

1. The data generating processes DGP-I and DGP-II are defined in equations (4) and (5). In these size experiments, the coefficient b in each DGP is set to 0. In each simulation, 1–step ahead forecasts of y are formed from an estimated AR model for y and an estimated VAR in y and x.
2. In each simulation, the lag lengths of the estimated models are set at the true lag order of the DGP.
3. R and P refer to the number of in–sample observations and post–sample predictions, respectively.
4. Sections 3 and 4 in the text describe the test statistics. In the results based on asymptotic critical values for $\pi = \hat{\pi} \equiv P/R$, the statistics are compared against critical values taken from Table 1 and McCracken (1999). In the results based on standard normal critical values, the statistics are compared against the asymptotic distribution of the tests for $\pi = 0$.
5. The number of simulations is 50,000.

Table 3

Selected Results On Empirical Size When Estimated Model Lags Are Data-Determined

Recursive Forecasts

Nominal Size = 10%

	$R = 100,\ P = 20$		$R = 100,\ P = 40$		$R = 100,\ P = 100$		$R = 200,\ P = 20$		$R = 200,\ P = 40$		$R = 200,\ P = 200$	
	AIC	*SIC*	*AIC*	*SIC*	*AIC*	*SIC*	*AIC*	*SIC*	*AIC*	*SIC*	*AIC*	*SIC*
DGP-I												
Tests Compared Against Asymptotic Critical Values for $\pi = \hat{\pi} \equiv P/R$												
MSE-F	.110	.107	.102	.100	.103	.106	.110	.107	.107	.105	.100	.102
MSE-T	.128	.132	.116	.118	.102	.106	.130	.133	.118	.121	.101	.103
MSE-REG	.114	.117	.109	.110	.100	.104	.116	.119	.109	.111	.099	.101
ENC-NEW	.118	.110	.110	.103	.110	.105	.118	.111	.113	.106	.106	.101
ENC-T	.134	.135	.120	.118	.110	.110	.129	.131	.123	.123	.105	.104
ENC-REG	.119	.118	.108	.107	.105	.106	.115	.117	.113	.113	.102	.101
GC	.133	.102	.133	.103	.134	.104	.133	.103	.129	.099	.128	.101
Tests Compared Against Standard Normal Critical Values												
MSE-T	.058	.060	.039	.040	.018	.020	.071	.073	.050	.052	.017	.018
MSE-REG	.047	.049	.033	.034	.017	.018	.060	.062	.044	.045	.015	.016
ENC-T	.091	.090	.079	.076	.067	.065	.094	.093	.081	.079	.062	.060
ENC-REG	.078	.076	.071	.068	.062	.060	.083	.082	.073	.071	.060	.057
DGP-II												
Tests Compared Against Asymptotic Critical Values for $\pi = \hat{\pi} \equiv P/R$												
MSE-F	.145	.258	.144	.292	.159	.410	.111	.180	.112	.197	.103	.263
MSE-T	.144	.227	.143	.261	.153	.375	.123	.163	.114	.173	.102	.244
MSE-REG	.130	.207	.136	.252	.151	.370	.109	.148	.106	.163	.101	.242
ENC-NEW	.159	.299	.162	.344	.181	.470	.119	.205	.118	.225	.112	.296
ENC-T	.157	.264	.157	.309	.172	.443	.123	.176	.120	.200	.107	.282
ENC-REG	.142	.244	.146	.294	.168	.437	.110	.162	.111	.189	.105	.279
GC	.161	.339	.152	.331	.132	.251	.128	.217	.130	.192	.127	.105
Tests Compared Against Standard Normal Critical Values												
MSE-T	.066	.119	.048	.120	.040	.147	.068	.095	.048	.085	.014	.097
MSE-REG	.056	.103	.043	.109	.037	.140	.061	.085	.042	.078	.012	.095
ENC-T	.119	.198	.117	.240	.131	.362	.103	.143	.090	.154	.075	.237
ENC-REG	.106	.180	.108	.226	.126	.354	.092	.131	.082	.144	.072	.234

Notes:

1. The data generating processes DGP-I and DGP-II are defined in equations (4) and (5). In these size experiments, the coefficient b in each DGP is set to 0. In each simulation, 1–step ahead forecasts of y are formed from an estimated AR model for y and an estimated VAR in y and x.

2. The table reports size results based on setting the lag lengths of the models estimated in each simulation to minimize the AIC or SIC for the estimated VAR model.

3. R and P refer to the number of in–sample observations and post–sample predictions, respectively.

4. Sections 3 and 4 in the text describe the test statistics. In the results based on asymptotic critical values for $\pi = \hat{\pi} \equiv P/R$, the statistics are compared against critical values taken from Table 1 and McCracken (1999). In the results based on standard normal critical values, the statistics are compared against the asymptotic distribution of the tests for $\pi = 0$.

5. The number of simulations is 50,000.

Table 4

Size–Adjusted Power, $b = .1$

Recursive Forecasts

(Empirical Size = 10%)

	$R = 50$			$R = 100$					$R = 200$			
	P=100	P=150	P=250	P=10	P=20	P=40	P=100	P=200	P=20	P=40	P=80	P=200
DGP-I												
MSE-F	.339	.422	.567	.214	.239	.284	.379	.535	.299	.345	.419	.592
MSE-T	.320	.410	.566	.141	.177	.227	.338	.507	.179	.234	.320	.522
MSE-REG	.321	.411	.566	.147	.185	.232	.339	.507	.187	.241	.325	.523
ENC-NEW	.375	.454	.585	.234	.268	.332	.452	.607	.342	.414	.518	.702
ENC-T	.361	.447	.595	.153	.202	.267	.404	.586	.211	.289	.409	.641
ENC-REG	.367	.452	.598	.167	.210	.272	.407	.590	.222	.298	.414	.643
GC	.399	.477	.629	.324	.340	.380	.482	.630	.508	.546	.599	.744
DGP-II												
MSE-F	.281	.352	.459	.172	.193	.228	.305	.440	.242	.283	.337	.486
MSE-T	.275	.352	.470	.134	.153	.192	.282	.425	.165	.208	.272	.448
MSE-REG	.275	.352	.470	.138	.158	.194	.284	.427	.173	.212	.275	.447
ENC-NEW	.311	.385	.492	.188	.220	.264	.366	.515	.284	.345	.418	.603
ENC-T	.305	.385	.503	.145	.170	.223	.333	.494	.188	.251	.338	.547
ENC-REG	.307	.389	.503	.150	.178	.228	.337	.497	.198	.258	.341	.550
GC	.331	.410	.528	.254	.277	.310	.398	.541	.426	.466	.501	.654

Notes:

1. The data generating processes DGP-I and DGP-II are defined in equations (4) and (5). In these power experiments, the coefficient b in each DGP is set to .1. In each simulation, 1–step ahead forecasts of y are formed from an estimated AR model for y and an estimated VAR in y and x.
2. In each simulation, the lag lengths of the estimated models are set at the true lag order of the DGP.
3. R and P refer to the number of in–sample observations and post–sample predictions, respectively.
4. Sections 3 and 4 in the text describe the test statistics. In each experiment, power is calculated by comparing the test statistics against empirical critical values, calculated as the 90th percentile of the distributions of the statistics in the corresponding size experiment (in which the DGP, R, and P are the same as in the power experiment, except $b = 0$).
5. The number of simulations is 10,000.

Table 5

Size–Adjusted Power, $b = .2$

Recursive Forecasts

(Empirical Size = 10%)

	$R = 50$			$R = 100$					$R = 200$			
	P=100	P=150	P=250	P=10	P=20	P=40	P=100	P=200	P=20	P=40	P=80	P=200
DGP-I												
MSE-F	.747	.853	.958	.411	.483	.589	.777	.927	.553	.649	.773	.932
MSE-T	.707	.834	.958	.211	.294	.431	.688	.896	.274	.407	.596	.871
MSE-REG	.711	.836	.958	.232	.313	.443	.692	.897	.296	.420	.602	.872
ENC-NEW	.841	.921	.980	.492	.599	.732	.908	.981	.693	.819	.927	.994
ENC-T	.815	.911	.982	.266	.389	.580	.850	.976	.388	.594	.810	.982
ENC-REG	.821	.915	.983	.301	.419	.597	.854	.976	.423	.612	.822	.984
GC	.863	.934	.987	.749	.780	.839	.935	.988	.954	.967	.984	.998
DGP-II												
MSE-F	.667	.789	.924	.340	.413	.506	.706	.888	.481	.587	.700	.902
MSE-T	.649	.789	.933	.195	.270	.387	.643	.872	.264	.385	.554	.856
MSE-REG	.651	.789	.932	.209	.280	.395	.646	.874	.279	.397	.563	.856
ENC-NEW	.763	.865	.960	.408	.516	.642	.856	.966	.621	.761	.880	.988
ENC-T	.740	.856	.962	.242	.349	.513	.792	.953	.363	.547	.755	.967
ENC-REG	.746	.861	.963	.266	.372	.527	.798	.955	.388	.572	.766	.968
GC	.798	.891	.976	.663	.699	.766	.888	.976	.920	.941	.963	.995

Notes:

1. The data generating processes DGP-I and DGP-II are defined in equations (4) and (5). In these power experiments, the coefficient b in each DGP is set to .2. In each simulation, 1–step ahead forecasts of y are formed from an estimated AR model for y and an estimated VAR in y and x.
2. In each simulation, the lag lengths of the estimated models are set at the true lag order of the DGP.
3. R and P refer to the number of in–sample observations and post–sample predictions, respectively.
4. Sections 3 and 4 in the text describe the test statistics. In each experiment, power is calculated by comparing the test statistics against empirical critical values, calculated as the 90th percentile of the distributions of the statistics in the corresponding size experiment (in which the DGP, R, and P are the same as in the power experiment, except $b = 0$).
5. The number of simulations is 10,000.

Table 6

Testing the Predictive Content of Unemployment for Inflation

Recursive Forecasts

$R = 115,\ P = 46$

In–Sample Model Estimates

Explanatory	*Dependent variable*		
variable	$\Delta Inflation_t$	$\Delta Inflation_t$	$\Delta Unemployment_t$
Constant	.024 (.154)	.033 (.148)	-.009 (.031)
$\Delta Inflation_{t-1}$	-.288 (.092)	-.391 (.093)	.057 (.019)
$\Delta Inflation_{t-2}$	-.237 (.092)	-.266 (.097)	.015 (.020)
$\Delta Unemployment_{t-1}$		-1.207 (.454)	.703 (.093)
$\Delta Unemployment_{t-2}$		-.137 (.457)	-.182 (.094)
$\bar{R}^2$	.092	.166	.356

Tests of Predictive Power of Unemployment for Inflation

	Test	*Asymptotic*	*Empirical*
	statistics	*critical values*	*critical values*
		for $\pi = .4$	
MSE, AR	.420		
MSE, VAR	.412		
MSE-F	.839	1.029	1.110
MSE-T	.099	.614	.701
MSE-REG	.137	.614	.666
ENC-NEW	5.186	1.019	1.079
ENC-T	1.112	1.086	1.178
ENC-REG	1.698	1.086	1.139
GC	8.107	2.337	2.474

Notes:
1. The figures in parentheses in the upper panel of the table are standard errors for the reported coefficient estimates.
2. 1–step ahead forecasts of the change in inflation are formed from an estimated AR model for the change in inflation and an estimated VAR in the changes in inflation and unemployment.
3. R and P refer to the number of in–sample observations and post–sample predictions, respectively. The in–sample and post–sample periods span 1958:Q3 to 1987:Q1 and 1987:Q2 to 1998:Q3.
4. The significance level of the tests is 10%.
5. Sections 3 and 4 in the text describe the test statistics. The asymptotic critical values are taken from Table 1 and McCracken (1999).
6. The empirical critical values are generated from a Monte Carlo experiment (using 50,000 simulations) in which the DGP is a VAR in the changes in inflation and unemployment imposing the null that unemployment not enter the inflation equation. The equations of the simulated model, estimated with just in–sample data, are given in columns 2 and 4 of the top panel. The covariance matrix of the residuals in the DGP is

$$\text{Var}\begin{pmatrix} u_{infl,t} \\ u_{unemp,t} \end{pmatrix} = \begin{pmatrix} 2.673 & -.081 \\ -.081 & .102 \end{pmatrix}.$$