Inference About Predictive Ability

*Michael W. McCracken and
Kenneth D. West*

14.1. INTRODUCTION

Traditionally, statistical evaluation of an econometric model focuses on "in-sample" analysis of the residuals from a fitted model. This methodology has powerful theoretical and practical justification. However, it is not always particularly natural or effective. A number of studies (see references below) find that models that seem to fit well by conventional in-sample criteria do poorly at out-of-sample prediction. This has led observers such as Klein (1992) to argue that the "ability to make useful ex-ante forecasts is the *real* test of a model."

In this chapter we provide a brief review of how out-of-sample methods can be used to construct tests that evaluate a time-series model's ability to predict. Although we touch on a wide range of issues, we focus on the role that parameter estimation plays in constructing asymptotically valid tests of predictive ability. We explain why forecasts and forecast errors that depend upon estimated parameters do not necessarily have the statistical properties of their population counterparts: asymptotic inference sometimes requires explicitly accounting for uncertainty about the population values of such parameters. Although this is well known when using fitted values and residuals (for example, Durbin, 1970; Pagan and Hall, 1983), it is commonly ignored when using out-of-sample methods.

We survey a literature that evaluates forecasts and forecast errors generated over an "out-of-sample" period. The forecasts may be truly *ex ante*, generated before the data on the predictand are available. More commonly, the investigator sets the out-of-sample period aside, specifically to allow model evaluation. Dawid (1984), who refers to it as the "prequential" approach to testing, has paid considerable attention to this method.

Because the sample is split into in-sample and out-of-sample portions, this method of evaluating a model is somewhat different from that commonly taught in econometric textbooks, in both minor and major ways. An example of a technical difference is that in many applications parameters are updated as more data becomes available throughout the out-of-sample period; if so, each new forecast is constructed using a potentially different estimate of the population level parameter vector. More generally, since forecasts of (say) y_{t+1} may lie off of the time t empirical support, we expect the predictions to be less prone to biases induced by parameter estimation and, more generally, to overfitting and pre-test biases.

The remainder of the chapter proceeds as follows. Section 14.2 provides a review of the literature in which out-of-sample methods have been used to test for predictive ability. Section 14.3 consists of four subsections. The first is an analytic example illustrating why error in estimation of regression parameters used to make forecasts affects the asymptotic distribution of test statistics. The second and third subsections present general discussions of the asymptotic effects of such error, and of how to feasibly account for such error in construction of test statistics. The final subsection reviews Monte Carlo evidence on the importance of such accounting. Section 14.4 concludes and provides directions for future research.

One bit of terminology: because we need to refer frequently to error induced by estimation of regression parameters used to make forecasts, we denote such error as "parameter estimation error" or "parameter uncertainty." Similarly, "parameter estimation" is used as shorthand for "estimation of regression parameters used to make forecasts," an abuse of terminology in that the out-of-sample object of interest (say, mean squared forecast error) is also a parameter that is estimated.

14.2. Literature Review

While use of out-of-sample forecasts for model evaluation is less common than use of in-sample evidence, forecast evaluation has a long and distinguished history in economics. Early references include Christ (1956) and Goldberger (1959), each of which evaluates the predictive ability of the Klein–Goldberger model of the U.S. economy. Here we review more recent papers, citing both methodological contributions and recent representative applied papers.

There are many measures of forecast quality, but most tests that are used in practice can be grouped into one of five categories: *equal forecast accuracy* between two (or more) predictive models, forecast *encompassing*, forecast *efficiency*, zero forecast *bias*, and *sign predictability*. We discuss each in turn while referencing relevant methodological and applied work. Throughout, we assume that the forecasting model is parametric and, for expositional ease, that the (scalar) predictand is y_{t+1}. The parameter estimate $\hat{\beta}_t$ is a function of observables known at time t. Let the forecast be based upon a function $\hat{y}_{t+1}(\beta)$ such that the time $t = R, \ldots, T - 1$ forecast of y_{t+1} is $\hat{y}_{t+1} \equiv \hat{y}_{t+1}(\hat{\beta}_t)$. If we let P denote the number of one-step-ahead forecast errors $\hat{u}_{t+1} = y_{t+1} - \hat{y}_{t+1}$, then

$$R + P = T + 1. \tag{14.1}$$

When two forecasting models exist, an additional index $i = 1, 2$ will be used to distinguish between the parameter estimates, forecasts, and forecast errors from each of the two models.

A very general test of predictive ability is one that tests for *equal forecast accuracy* across two or more models. To construct a test of this type one needs to select a measure of accuracy. McCulloch and Rossi (1990), Leitch and Tanner (1991), and West, Edison and Cho (1993) compare predictive ability using economic measures of accuracy. It is more common to use statistical measures. In particular, the most common comparison is whether two predictive models have the same mean squared error (MSE).

Granger and Newbold (1977) propose a test for equal MSE that is similar to a test suggested by Morgan (1939). They note that if the sequence of forecast errors form a random sample from a bivariate normal population then the sample correlation coefficient, associated with $(\hat{u}_{1,t+1} + \hat{u}_{2,t+1})$ and $(\hat{u}_{1,t+1} - \hat{u}_{2,t+1})$, can be used to construct a test for equal MSE. This test has been applied by a number of authors, including Granger and Deutsch (1992), who compare the predictive ability of models of the unemployment rate. Ashley, Granger, and Schmalensee (1980), Ashley (1981), Alpert and Guerard (1988), and Bradshaw and Orden (1990) use the statistic to test for causality between two nested models.

Extensions include Brandt and Bessler (1983), who explicitly adjust for the fact that forecast errors may be biased away from zero. A test used by Meese and Rogoff (1988) and Diebold and Rudebusch (1991) allows for serial correlation in the forecast errors. Park (1990), Tegene and Kuchler (1994), and Granger and Swanson (1997) apply a variant of this statistic based upon Fisher's z-test.

A more general test for equal forecast accuracy is that suggested by Diebold and Mariano (1995). They base their test for equal MSE upon the loss differential, $d_{t+1} = \hat{u}_{1,t+1}^2 - \hat{u}_{2,t+1}^2$. Let $\gamma_j = E d_{t+1} d_{t+1-j}$ and let $\hat{S}_{dd}$ denote a consistent estimate of $S_{dd} = \gamma_0 + 2\sum_{j=1}^{\infty} \gamma_j$, the long-run covariance associated with the covariance stationary sequence d_{t+1}. The authors argue that the statistic $P^{-1/2} \sum_{t=R}^{R+P-1} d_{t+1} / \hat{S}_{dd}^{1/2}$ converges in distribution to a standard normal variable and hence normal tables can be used to test the null of equal forecast accuracy.

This statistic has been frequently used in recent years. Engel (1994), Chinn and Meese (1995), and Blomberg and Hess (1997) test whether regime-shifting models, error correction models, and models of political behavior, respectively, outperform the random walk in the prediction of exchange rates. Mark (1995) and Kilian (1999) use a bootstrapped version of this statistic to compare the predictive ability of long-horizon models of exchange rates to the random walk. Bram and Ludvigson (1998) use a modified version of this test, suggested by Harvey, Leybourne, and Newbold (1997), to determine a causal relationship between consumer confidence and household expenditure.

An advantage of the technique proposed by Diebold and Mariano (1995) is that it can be extended to other measures of loss (say) $L(.)$. To do so we need only define $d_{t+1} = L(\hat{u}_{1,t+1}) - L(\hat{u}_{2,t+1})$ and proceed to construct the test as we did when the loss was quadratic. Based upon this observation, Chen and Swanson (1996)

use squared forecast error, absolute forecast error, and absolute percentage forecast error as loss functions to evaluate several parametric and semiparametric forecasting models of U.S. inflation. Swanson and White (1997a,b) use the same three loss functions to evaluate how well parametric and neural net models predict several macroeconomic variables.

The test of forecast *encompassing* is related to the literature on the combination of forecasts introduced by Bates and Granger (1969) (see Diebold, 1989). The idea of forecast combination is that if two (or more) forecasting models are available, then taking a weighted combination of the available forecasts may generate a better forecast. Granger and Ramanathan (1984) suggest doing this using regression-based methods. Suppose that $\hat{y}_{1,t+1}$ and $\hat{y}_{2,t+1}$ are competing forecasts for y_{t+1} and that we already have an observed series of these *ex ante* forecasts. They construct future forecasts using the weights α_0, α_1, and α_2 estimated by the OLS regression

$$y_{t+1} = \alpha_0 + \alpha_1 \hat{y}_{1,t+1} + \alpha_2 \hat{y}_{2,t+1} + \text{error term.} \qquad (14.2)$$

Based upon this regression, Chong and Hendry (1986) observe that under the null that forecasts from model 1 encompass the forecasts from models 2, α_1 and α_2 should be 1 and 0 respectively. If this is the case, then a test for forecast encompassing can be constructed using the t-statistic associated with α_2 in the OLS estimated regression

$$\hat{u}_{1,t+1} = \alpha_0 + \alpha_2 \hat{y}_{2,t+1} + \text{error term.} \qquad (14.3)$$

Under the null that the model 1 forecast encompasses model 2, they argue that the t-statistic should be asymptotically standard normal and hence normal tables can be used to conduct a test of encompassing.

Tests related to that in (14.3) include one suggested by Fair and Shiller (1989, 1990). They construct a regression-based test that is closer to (14.2) than (14.3). Their test for "information content" in m-step-ahead forecasts does not impose the null that one particular model forecast encompasses the other and hence can be considered more general than that in (14.3). Ericsson (1992) constructs a regression-based test that is designed to allow for forecasts generated by models with cointegrating relationships. He notes that if the population level forecast errors are I(0) but the population level forecasts are I(1), then equation (14.3) is "unbalanced." To rectify this possibility he suggests using the t-statistic associated with $(\hat{y}_{2,t+1} - \hat{y}_{1,t+1})$ when it and a constant are regressed on $\hat{u}_{1,t+1}$.

These tests of encompassing have been implemented by a number of authors. Day and Lewis (1992) determine the information content in volatility forecasts. Ericsson and Marquez (1993) test whether certain models of the U.S. trade balance forecast encompass other models. They also argue that by estimating the equation using GLS they can account for the "asymptotically negligible" autocorrelation in the forecast errors due to parameter uncertainty. Oliner, Rudebusch, and Sichel (1995) determine whether "new" models of business investment forecast encompass "old" models. Lo and MacKinlay (1997) test for forecast encompassing in

models of asset returns from the stock and bond markets. Amano and van Norden (1995) test whether or not error correction models forecast encompass the random walk model. Andrews, Minford, and Riley (1996) conduct a series of tests of encompassing that includes (14.3). Fair (1999) compares the information content in structural models of exchange rates with the random walk model. Stock and Watson (1999a,b) test for information content in forecasts of inflation and several other macroeconomic variables.

Mincer and Zarnowitz (1969) introduce the test of forecast *efficiency*. The test for forecast efficiency is based upon the observation that if the forecast is constructed using all available information, then the optimal forecast (in the sense of minimum mean square error) and the forecast error should be uncorrelated. If this is the case, then a test of efficiency may be constructed using the t-statistic associated with the estimate of α_1 in the OLS estimated regression

$$\hat{u}_{t+1} = \alpha_0 + \alpha_1 \hat{y}_{t+1} + \text{error term.} \tag{14.4}$$

Authors who have used this test of efficiency include the following. McNees (1978) estimates (14.4) by both OLS and GLS to allow for multi-step forecasts and then applies the method to predictions of the unemployment rate, real GNP, and the GNP deflator. Berger and Krane (1985) evaluate the efficiency of forecasts of real GNP made by D.R.I. and Chase Econometrics. Pagan and Schwert (1990), Day and Lewis (1992), and West and Cho (1995) evaluate the efficiency of volatility forecasts. Stock and Watson (1993) construct a comparable test for the efficiency of forecasts of the probability of recessions. Hatanaka (1974) and Keane and Runkle (1989) have suggested extensions to this type of test.

The test for zero forecast *bias* is sometimes referred to as a test for *zero-mean prediction error*. Mincer and Zarnowitz (1969) introduce this test in the context of tests of efficiency. For example, note that in (14.4) we have included an intercept term in the regression. They note that if the forecasts are unbiased then the intercept term should be zero. Ericsson and Marquez (1993) also give this argument in the context of regression-based tests of encompassing like that in (14.3). Because of this, tests for zero bias are commonly reported along with (and tested jointly with) tests of efficiency and encompassing. Day and Lewis (1992) report the t-statistics associated with the intercept term in both regression-based tests of efficiency as in (14.4) and regression-based tests of encompassing as in (14.3). Berger and Krane (1985) use an F-test to construct a joint test for both efficiency and zero bias. Ericsson and Marquez (1993) do the same but for encompassing and zero bias. It is also possible to take a more direct approach when testing for zero bias. Stock and Watson (1993) and Oliner, Rudebusch, and Sichel (1995) construct a test for zero bias by regressing the forecast error on a constant. They then use the t-statistic associated with the intercept term to test for zero bias.

Perhaps the most well known test of *sign predictability* was developed in a series of papers by Merton (1981) and Henriksson and Merton (1981). These authors are interested in evaluating whether decisions based upon $\hat{y}_{t+1}$ are useful in the absence of knowing the actual value of y_{t+1}. The context that the authors had in mind was one where decisions had to be made to either buy or sell an asset.

For example, let y_{t+1} denote the return on an asset and let $\hat{y}_{t+1}$ denote the forecasted return. If the sign of $\hat{y}_{t+1}$, denoted sgn $(\hat{y}_{t+1})$, is positive then the decision is made to buy shares of the asset. If sgn $(\hat{y}_{t+1})$ is negative then the decision is made to sell shares of the asset. Notice that the actual value of the forecast is not important *per se*, just the sign of the forecast.

In this decision-based forecasting literature, we are interested in whether the sign of $\hat{y}_{t+1}$ is useful as a predictor of the sign of y_{t+1}. If it is not useful then sgn $(\hat{y}_{t+1})$ is independent of sgn (y_{t+1}). To test this null hypothesis, Henriksson and Merton (1981) suggest a test using the t-statistic associated with α_1 in the following OLS estimated regression:

$$1(\hat{y}_{t+1} \geq 0) = \alpha_0 + \alpha_1 1(y_{t+1} \geq 0) + \text{error term}, \tag{14.5}$$

where the function $1(.)$ takes the value one if the argument is true and zero otherwise.

Applications of this test of sign predictability include Breen, Glosten, and Jagannathan (1989), who are concerned with the sign predictability of indices of stock returns. Park (1990) tests for sign predictability of changes in the price of U.S. cattle. Tegene and Kuchler (1994) evaluate the sign predictive ability of changes in farmland prices in the U.S. Midwest. Kuan and Liu (1995), Chen and Swanson (1996), and Swanson and White (1995, 1997a,b) evaluate the sign predictive ability of both parametric and neural net models.

Other sign tests include one suggested by Cumby and Modest (1986), who argue that their test of sign predictability is more powerful than that in (14.5). Pesaran and Timmermann (1992) suggest a test, similar to that in (14.5), that they use in Pesaran and Timmermann (1995). More recently, Pesaran and Timmermann (1994) extend the Henriksson and Merton (1981) test to situations where there are more than two choices available. This is potentially useful since, for example, when making investment decisions one usually has the choice to not only buy or sell an asset but also to hold and not act in the relevant market. Chinn and Meese (1995) use a test of sign predictability that is based upon the binomial distribution.

It is interesting to note that this literature on sign prediction runs parallel to the literature on probability forecasting that is particularly important in meteorology. For example, suppose that a decision is made to take an action based upon the probability forecast, p_{t+1}, of rain. A farmer has to make a decision as to whether to spread fertilizer onto his fields. If it rains, then the fertilizer may wash away and not have time to penetrate the soil. In this case, the farmer makes the decision to spread the fertilizer only if the probability forecast of rain is (say) less than 20 percent. In this case, the farmer is interested in the sign of $\hat{y}_{t+1} \equiv p_{t+1} - 20\%$. The obvious difficulty with evaluating these types of forecasts is that we do not have the luxury of observing the actual probability p_{t+1} unless $p_{t+1} = 1.00$ or $p_{t+1} = 0$. See Brier (1950), as well as the review by Murphy and Daan (1985), for a discussion on the evaluation of probability forecasts in meteorology. See Granger and Pesaran (2000) and Lopez (2001) for discussions relevant to economists.

One common feature of all of the *applied* papers cited above is that the predictions are made using estimated regression models. This is in sharp contrast to all

of the *methodological* papers cited above. None of these explicitly consider possible effects from use of an estimated regression model to construct forecasts. That is, their arguments make use of conditions that are not obviously applicable when such models are used.

This difference, between the applied papers and the methodological papers, raises a question. Do the results of the methodological papers remain valid when regression models are estimated rather than known? West (1996) and McCracken (2000) use first-order asymptotics to address this question, assuming parametric regression models and stationary data. Their results delineate conditions under which the results of these papers remain valid, and, more generally, present the appropriate limiting results when forecasts rely on estimated regression parameters. A discussion of these and other theoretical results, along with some related Monte Carlo evidence, is provided in the next section.

14.3. THEORETICAL AND MONTE CARLO EVIDENCE

This section contains four subsections. The first illustrates how parameter estimation error can affect the asymptotic distribution of out-of-sample test statistics. The second section presents general asymptotic results, focusing on the results in West (1996) and McCracken (2000). The third section describes how to feasibly account for parameter estimation error. The final section summarizes some Monte Carlo results.

14.3.1. How parameter estimation affects inference: An example

Before presenting the theoretical results, we provide a simple illustrative example of how parameter estimation error can affect the limiting distribution of out-of-sample test statistics. Suppose that the forecasting model is a simple univariate linear regression estimated by OLS. If we let y_t denote the scalar predictand and x_{t-1} denote the scalar covariate, then we have the regression model

$$y_t = \beta_0^* + x_{t-1}\beta_1^* + u_t = X_{t-1}'\beta^* + u_t, \tag{14.6}$$

where $X_{t-1} = (1, x_{t-1})'$ and $\beta^* = (\beta_0^*, \beta_1^*)'$. For the sake of presentation let u_t be an i.i.d. zero mean conditionally homoskedastic disturbance term with unconditional variance σ^2 and let X_{t-1} be covariance stationary. Suppose that the one-step-ahead predictions are made in the obvious way using $\hat{y}_{t+1} = X_t'\hat{\beta}_t$. Write the one-step-ahead forecast error as $\hat{u}_{t+1} = y_{t+1} - \hat{y}_{t+1}$. For ease of presentation, assume that the parameter estimate $\hat{\beta}_t$ is estimated once by OLS using information available through $s = 1, \ldots, R$. Hence for any $t \geq R$, $\hat{\beta}_t = (R^{-1}\sum_{s=1}^{R}X_{s-1}X_{s-1}')^{-1}(R^{-1}\sum_{s=1}^{R}X_{s-1}y_s)$.

Consider the test for zero-mean prediction error conducted by, for example, Oliner, Rudebusch, and Sichel (1995). The test for zero-mean prediction error can be constructed using the t-statistic associated with α from the regression $\hat{u}_{t+1} = \alpha +$ error term. The t-statistic has the form

$$P^{-1/2} \sum_{t=R}^{R+P-1} \hat{u}_{t+1} / [(P-1)^{-1} \sum_{t=R}^{R+P-1} (\hat{u}_{t+1} - \bar{u})^2]^{1/2}, \tag{14.7}$$

where $\bar{u}$ is the sample average of the forecast errors $\hat{u}_{t+1}$.

To show that a statistic of this type is limiting standard normal usually requires three steps. The first is to show that the numerator of the t-statistic is asymptotically normal with a limiting variance Ω. The second is to show that the denominator converges in probability to $\Omega^{1/2}$. The final step is simply to apply Slutsky's Theorem and conclude that the t-statistic is asymptotically standard normal.

Consider the numerator. By definition,

$$P^{-1/2} \sum_{t=R}^{R+P-1} \hat{u}_{t+1} \tag{14.8}$$

$$= P^{-1/2} \sum_{t=R}^{R+P-1} u_{t+1} + (-P^{-1} \sum_{t=R}^{R+P-1} X_t')(R^{-1} \sum_{s=1}^{R} X_{s-1} X_{s-1}')^{-1} ((P/R)^{1/2} R^{-1/2} \sum_{s=1}^{R} X_{s-1} u_s).$$

If we assume, as in Hoffman and Pagan (1989), that

$$P, R \to \infty, \ P/R \to \pi < \infty, \tag{14.9}$$

then it is straightforward to show that, under general conditions,

$$P^{-1/2} \sum_{t=R}^{R+P-1} \hat{u}_{t+1} = (1 - (EX_t')(EX_t X_t')^{-1}) \begin{pmatrix} P^{-1/2} \sum_{t=R}^{R+P-1} u_{t+1} \\ \pi^{1/2} R^{-1/2} \sum_{s=1}^{R} X_{s-1} u_s \end{pmatrix} + o_p(1). \tag{14.10}$$

Since the disturbance terms are i.i.d., zero mean, and conditionally homoskedastic, we know that the two components of the column vector in (14.10) are independent and hence

$$\begin{pmatrix} P^{-1/2} \sum_{t=R}^{R+P-1} u_{t+1} \\ \pi^{1/2} R^{-1/2} \sum_{s=1}^{R} X_{s-1} u_s \end{pmatrix} \to_d N \left(0_{3 \times 1}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \pi \sigma^2 (EX_t X_t') \end{pmatrix}^{3 \times 3} \right). \tag{14.11}$$

We then know that $P^{-1/2} \sum_{t=R}^{R+P-1} \hat{u}_{t+1} \to_d N(0, \Omega)$, where

$$\Omega = \sigma^2 + \pi \sigma^2 (EX_t')(EX_t X_t')^{-1}(EX_t). \tag{14.12}$$

The right-hand side of (14.12) should seem familiar. It is directly comparable to the variance of a one-step-ahead prediction error (Goldberger, 1991, p. 175). It differs because it is the variance of a scaled sample average of one-step-ahead predictions rather than just a single one-step-ahead prediction error.

Now consider the denominator in (14.7). To show that $(P-1)^{-1} \sum_{t=R}^{R+P-1} (\hat{u}_{t+1} - \bar{u})^2 = (P-1)^{-1} \sum_{t=R}^{R+P-1} \hat{u}_{t+1}^2 - P(\bar{u})^2/(P-1)$ converges in probability to σ^2, first notice that equations (14.8)–(14.12) imply that $P(\bar{u})^2/(P-1)$ converges in probability to zero.

We then need only show that $(P-1)^{-1}\sum_{t=R}^{R+P-1}\hat{u}_{t+1}^2$ converges in probability to σ^2. By definition,

$$(P-1)^{-1}\sum_{t=R}^{R+P-1}\hat{u}_{t+1}^2 = (P-1)^{-1}\sum_{t=R}^{R+P-1}u_{t+1}^2 - 2((P-1)^{-1}\sum_{t=R}^{R+P-1}u_{t+1}X_t')(\hat{\beta}_R - \beta^*)$$

$$+ (\hat{\beta}_R - \beta^*)'((P-1)^{-1}\sum_{t=R}^{R+P-1}X_tX_t')(\hat{\beta}_R - \beta^*). \qquad (14.13)$$

Since the first right-hand side term in (14.13) converges in probability to σ^2, the result will follow if the last two right-hand side terms converge in probability to zero. This is evident, since $(P-1)^{-1}\sum_{t=R}^{R+P-1}u_{t+1}X_t' \to_p 0$, $(P-1)^{-1}\sum_{t=R}^{R+P-1}X_tX_t' \to_p EX_tX_t'$, and $\hat{\beta}_R - \beta^* \to_p 0$.

By Slutsky's Theorem we can now conclude that the t-statistic in (14.7) converges in distribution to a normal variable with limiting variance V,

$$V = 1 + \pi(EX_t')(EX_tX_t')^{-1}(EX_t). \qquad (14.14)$$

This limiting distribution is in direct contrast to that when the parameters are known. If the parameters are known, then the t-statistic is asymptotically normal but with a limiting variance $V = 1$. Since parameters are generally not known, and since $\pi(EX_t')(EX_tX_t')^{-1}(EX_t)$ is positive semidefinite, (14.14) will be greater than or equal to 1. When (14.14) is greater than 1, any test that uses standard normal tables, without accounting for the extra term induced by parameter estimation, will reject the null of zero-mean prediction error too often and may lead to a less powerful test of the null.

14.3.2. How parameter estimation affects inference: A general result

The argument of the preceding section can be expanded to include parametric functions other than that used to construct the test of zero-mean prediction error. Suppose that we are interested in testing the (scalar) null that for some parametric function $f_{t+1}(\beta)$, $Ef_{t+1}(\beta^*) = \theta$. Here, β^* is the unknown vector of parameters needed to make a forecast. For one-step-ahead forecasts, examples of $f(.)$ include (in each case, $\theta = 0$):

Test	$f_{t+1}(\beta)$	(14.15)
Equal mean square error	$(y_{t+1} - X_{1,t}'\beta_1)^2 - (y_{t+1} - X_{2,t}'\beta_2)^2$	
Equal mean absolute error	$\|y_{t+1} - X_{1,t}'\beta_1\| - \|y_{t+1} - X_{2,t}'\beta_2\|$	
Forecast encompassing in (14.3)	$(y_{t+1} - X_{1,t}'\beta_1)X_{2,t}'\beta_2$	
Forecast efficiency in (14.4)	$(y_{t+1} - X_t'\beta)X_t'\beta$	
Zero-mean prediction error	$y_{t+1} - X_t'\beta$	
Sign predictability in (14.5)	$[1(X_t'\beta \geq 0) - E1(X_t'\beta \geq 0)][1(y_{t+1} \geq 0) - E1(y_{t+1} \geq 0)]$	

In (14.15), we have assumed that when two forecasting models are needed, they both are linear with covariate vector $X_{i,t}$ and parameter vector β_i^* for models $i = 1, 2$ respectively. We then define $\beta^* = (\beta_1^{*\prime}, \beta_2^{*\prime})'$.

West's (1996) results cover those functions in (14.15) associated with the tests of equal mean square error, encompassing, efficiency, and zero-mean prediction error. More generally, his results cover tests that use any twice continuously differentiable function $f(.)$. McCracken (2000) extends West's results to situations where the moment function $Ef(.)$ is continuously differentiable but where the function $f(.)$ need not be differentiable. This extension allows for the remaining tests in (14.15), those for equal mean absolute error and sign predictability. The key to both of their results is an asymptotic expansion like that in (14.10) that has two distinct components: a first component that involves the population level forecasts and forecast errors and a second that is due to the fact that parameters are not known and must be estimated.

Again suppose that we are interested in the (scalar) null $Ef_{t+1}(\beta^*) = \theta$. To test this null we first construct a scaled sample average of the form $P^{-1/2}\sum_{t=R}^{R+P-1}(f_{t+1}(\hat{\beta}_t) - \theta)$. Furthermore, let the sequence of parameter estimates be constructed using one of the three schemes that figure prominently in the forecasting literature: the *recursive, rolling,* or *fixed* schemes. These schemes are defined as follows, with the formulas applicable for OLS estimation given in (14.16) below.

The *recursive* scheme is used in, for example, Pagan and Schwert (1990). In this scheme, the parameter estimates are updated as more data becomes available. Hence for any time $t = R, \ldots, T - 1$ the parameter estimates use all relevant information from time $s = 1, \ldots, t$.

The *rolling* scheme is used in, for example, Swanson (1998). Here, the parameter estimates are updated as more data becomes available, but always using only the most recent (say) R observations. This scheme is sometimes used when there are concerns about changepoints and biases from the use of older observations.

Finally, the *fixed* scheme is used in, for example, Ashley, Granger, and Schmalensee (1980). Here the parameter estimate is estimated only once, using data from (say) 1 to R and hence is not updated as forecasting moves forward in time from $t = R, \ldots, T - 1$. This scheme may seem inefficient, but is frequently used in the neural net literature where computational concerns are a serious issue. This scheme was used for presentation purposes in the zero-mean prediction error example of (14.7).

To see how each of these parameter estimation schemes differ consider the simple linear regression in (14.6). If the parameter estimate(s) is constructed using OLS, then:

Scheme	Estimator	(14.16)
Recursive	$\hat{\beta}_t = (t^{-1}\sum_{s=1}^{t}X_{s-1}X'_{s-1})^{-1}(t^{-1}\sum_{s=1}^{t}X_{s-1}y_s)$,	(a)
Rolling	$\hat{\beta}_t = (R^{-1}\sum_{s=t-R+1}^{t}X_{s-1}X'_{s-1})^{-1}(R^{-1}\sum_{s=t-R+1}^{t}X_{s-1}y_s)$,	(b)
Fixed	$\hat{\beta}_t = (R^{-1}\sum_{s=1}^{R}X_{s-1}X'_{s-1})^{-1}(R^{-1}\sum_{s=1}^{R}X_{s-1}y_s)$.	(c)

Notice that each has essentially the same form but that they vary because different observations are being used.

To present our asymptotic results, suppose that we are considering a univariate least squares model. Asymptotic results extend directly when considering multivariate models, or two or more nonnested models, say to compare their forecast accuracy. They extend directly when the estimation method is maximum likelihood (ML) or generalized method of moments (GMM), as indicated in parenthetical comments in the exposition below. Finally, they extend to multi-step rather than one-step-ahead forecasts. Essential technical conditions for these extensions include stationarity (no unit roots), existence of moments of sufficient order, and, if two or more models are being compared, that the models be nonnested. See the cited papers for exact details.

In our least squares example, define $h_s = X_{s-1} u_s$ and let $H(t)$ equal $(t^{-1} \sum_{s=1}^{t} h_s)$, $(R^{-1} \sum_{s=t-R+1}^{t} h_s)$, and $(R^{-1} \sum_{s=1}^{R} h_s)$ respectively for the recursive, rolling, and fixed schemes. Let $B = (EX_t X_t')^{-1}$. (In general, h_s is the score if the estimation method is ML and is the set of orthogonality conditions used to estimate β^* if the estimation method is GMM; B is the inverse of the expectation of the Hessian (ML) or the asymptotic linear combination of orthogonality conditions (GMM).) Define

$$F = \partial E f_{t+1}(\beta^*) / \partial \beta. \tag{14.17}$$

For the test of zero-mean prediction error, for example, this is $-EX_t'$ (see (14.10)). Then

$$P^{-1/2} \sum_{t=R}^{R+P-1} (f_{t+1}(\hat{\beta}_t) - \theta) = P^{-1/2} \sum_{t=R}^{R+P-1} (f_{t+1}(\beta^*) - \theta) + FBP^{-1/2} \sum_{t=R}^{R+P-1} H(t) + o_p(1). \tag{14.18}$$

The first component involves the population level forecasts and forecast errors. The second arises because parameters are not known and must be estimated.

Define S as the long-run covariance of $(f_{t+1}(\beta^*) - \theta, h_t')'$ with diagonal elements S_{ff} and S_{hh} and off-diagonal element S_{fh}. That is,

$$S = \begin{pmatrix} S_{ff} & S_{fh} \\ S_{fh}' & S_{hh} \end{pmatrix}. \tag{14.19}$$

Under mild conditions, $P^{-1/2} \sum_{t=R}^{R+P-1} (f_{t+1}(\hat{\beta}_t) - \theta)$ converges in distribution to a normal variable with limiting covariance matrix Ω, defined by

$$\Omega = S_{ff} + \lambda_{fh}(FBS_{fh}' + S_{fh}B'F') + \lambda_{hh}FBS_{hh}B'F', \tag{14.20}$$

where

Scheme	λ_{fh}	λ_{hh}	(14.21)
Recursive	$1 - \pi^{-1} \ln (1 + \pi)$	$2[1 - \pi^{-1} \ln (1 + \pi)]$	
Rolling, $\pi \leq 1$	$\pi/2$	$\pi - \pi^2/3$	
Rolling, $1 < \pi < \infty$	$1 - (2\pi)^{-1}$	$1 - (3\pi)^{-1}$	
Fixed	0	π	

The formulae in (14.20) and (14.21) illustrate the statement made in the introduction that parameter estimation can affect the limiting variance of out-of-sample test statistics. In (14.20), the first component, S_{ff}, is precisely the limiting variance accounted for by authors such as Diebold and Mariano (1995). The remaining components arise from estimation of β^*.

The intuition for the form of the variance can be seen in (14.18). There are two sources of uncertainty. The first is the uncertainty that would exist even if population level parameters were known. This is represented by the term $P^{-1/2}\sum_{t=R}^{R+P-1}(f_{t+1}(\beta^*) - \theta)$. Authors such as Mincer and Zarnowitz (1969), Granger and Newbold (1977), Diebold and Mariano (1995), and Harvey, Leybourne, and Newbold (1997, 1998, 1999) have accounted for this uncertainty when deriving their out-of-sample tests of predictive ability.

The second type of uncertainty arises because parameters are not known and must be estimated. This is represented by the term $FBP^{-1/2}\sum_{t=R}^{R+P-1}H(t)$ in (14.18). This is the source of uncertainty not explicitly considered by the previously mentioned authors. It generates a variance term, $\lambda_{hh}FBS_{hh}B'F'$, and a covariance term, $\lambda_{fh}(FBS'_{fh} + S_{fh}B'F')$. Randles (1982) derives the in-sample equivalent of the expansion in (14.18) and the limiting variance in (14.20). He also discusses the importance of accounting for parameter estimation uncertainty when in-sample tests are constructed. As well, Hoffman and Pagan (1989) provide a similar result for out-of-sample inference when the fixed scheme is used.

We close this subsection by discussing several papers that derive additional asymptotic results. The first paper is White (2000), which develops a bootstrap test of predictive ability that accounts for potential effects due to parameter estimation. The bootstrap, which is also suggested by Ashley (1998), potentially brings asymptotic refinements. It is also a computationally convenient method to compare predictive ability across a potentially large number of models. Perhaps most importantly, White's bootstrap algorithm explicitly accounts for the potential existence of data snooping effects of the kind discussed by Lo and MacKinlay (1990).

Corradi, Swanson, and Olivetti (2001) extends some of the results of West (1996) and White (2000) to unit-root environments. One result concerns tests of equal forecast accuracy between two nonnested parametric regression models that contain cointegrating relationships. Based upon an expansion akin to that in (14.18) they show that when MSE is the measure of accuracy, parameter uncertainty is asymptotically irrelevant and hence the standard tests for equal forecast

accuracy will be asymptotically standard normal. When other measures of accuracy are used, they show that parameter uncertainty can have an effect. That effect causes the standard tests to have nonstandard limiting distributions that are not well approximated by the standard normal distribution. They also extend the results in White (2000) by validating the use of White's bootstrap method in environments where forecasts are generated using estimated regression models with cointegrating relationships.

McCracken (1999) and Clark and McCracken (2001) derive results for tests of equal forecast accuracy and encompassing when the two models are nested rather than nonnested. They show that these tests are not asymptotically normal but instead have nonstandard limiting distributions. This occurs since, when models are nested and the parameters are known in advance, the forecast errors associated with the two models are identical. In particular, the difference in (say) the loss differential d_{t+1} must be zero by definition. This set of results is potentially useful, since tests of causality by Ashley, Granger, and Schmalensee (1980), tests of long-horizon predictive ability by Mark (1995), and tests of market efficiency by Pesaran and Timmermann (1995) each involve the comparison of two nested rather than nonnested models.

Finally, Kitamura (1999) proposes bootstrap methods that can be used to improve the accuracy of point estimates of moments of functions of out-of-sample forecasts and forecast errors. In particular, he shows that the bootstrap can be used to smooth over the added variability in these point estimates induced by parameter estimation. His results are general enough to be applied to functions $f(.)$ that are either differentiable or nondifferentiable, and hence can be applied to measures of forecast accuracy like MSE and sign predictability.

14.3.3. Accounting for parameter uncertainty

We return now to the result in (14.20). We remarked in section 14.2 that the applied papers we cited do not explicitly account for error in estimation of parameters used to make forecasts. Even so, the analysis just presented can be used to show that, in some cases, the extra terms introduced by such error vanish, at least asymptotically. One important contribution of the preceding analysis is that it delineates an exact set of conditions under which there is no need to account for such error, thereby providing conditions under which the inference in those papers is asymptotically valid.

One such condition is that the out-of-sample size, P, is small relative to the in-sample size, R. This is transparent if we return to the limiting variance in (14.20). Each of the terms due to the existence of parameter estimation error depends upon the value π through λ_{fh} and λ_{hh}. In either case, if $\pi = 0$ then parameter estimation uncertainty is asymptotically irrelevant, since both $\lambda_{fh} = 0$ and $\lambda_{hh} = 0$ (the result for the recursive scheme follows, since $1 - \pi^{-1} \ln (1 + \pi) \to 0$ as $\pi \to 0$). When this is the case $\Omega = S_{ff}$ and hence parameter uncertainty can be safely ignored. As noted by Chong and Hendry (1986), the intuition behind this result is that when R is large relative to P, uncertainty about β^* is small relative to uncertainty that would be present even if β^* were known. Examples where this

condition may be useful include Fair (1980, $P/R = 0.156$), Pagan and Schwert (1990, $P/R = 0.132$ and 0.4), and Kuan and Liu (1995, $P/R = 0.042$, 0.087 and 0.137).

It is important to point out that it will not always, or perhaps even usually, be the case that P/R (the obvious sample analog to π) is small enough for parameter estimation error to be irrelevant for inference. There are many applications in which P/R is numerically large. Examples include Mark (1995), for which P/R ranges from 0.69 to 1.11, and Pesaran and Timmermann (1995), for which P/R ranges from 0.38 to 5.55. In any case, inspection of the asymptotic variance formula (14.20) shows that the value of π alone does not determine the effects of parameter estimation error; the parameters of the data generating process, and the estimator used, are also relevant. The Monte Carlo simulations summarized in section 14.3.4 show that serious biases result when parameter estimation error is ignored, for data generating processes, and choice of P and R, that are calibrated to actual economic data.

A second set of circumstances is also immediate from (14.20). When $F = 0$, the second and third terms disappear and hence $\Omega = S_{ff}$. Perhaps the most useful application of this result is to the test of equal MSE. When the test of interest is one of equal MSE between two linear parametric models, $F = (-2Eu_{1,t+1}X'_{1,t}, 2Eu_{2,t+1}X'_{2,t})$. Note that both components of F are essentially the moment conditions used to identify the parameters when they are estimated by OLS. Since these are both equal to zero we know that when OLS provides consistent estimates of the parameters it must be the case that F is a vector of zeroes. When this is the relevant environment, the Granger and Newbold (1977) and Diebold and Mariano (1995) tests of equal MSE can be applied without adjustment and still obtain a statistic that is asymptotically standard normal.

A third set of circumstances is not quite so obvious. We see in (14.21) that when the recursive scheme is used, $\lambda_{hh} = 2\lambda_{fh}$ and hence $\Omega = S_{ff} + \lambda_{fh}(FBS'_{fh} + S_{fh}B'F' + 2FBS_{hh}B'F')$. This is potentially useful since it is possible that

$$FBS'_{fh} + S_{fh}B'F' + 2FBS_{hh}B'F' = 0, \qquad (14.22)$$

regardless of whether F or π equals zero.

A particularly intriguing example of this occurs when one is testing for first-order serial correlation, disturbances are i.i.d., and a lagged dependent variable is used as a predictor in a linear regression model. From Durbin (1970) we know that when constructing an in-sample test for serial correlation this is potentially a problem. When the test is constructed out-of-sample, *and* the recursive scheme is used, this problem does not occur. The reason for this is that for this scalar test, it can be shown that $-FBS'_{fh} = FBS_{hh}B'F'$. This implies that (14.22) is satisfied and hence parameter estimation error is asymptotically irrelevant. For a discussion and other examples of this type of cancellation, see West (1996) and West and McCracken (1998).

Note that for the rolling and fixed schemes, parameter estimation error is asymptotically relevant for this test (because for these schemes, $\lambda_{hh} \neq 2\lambda_{fh}$). When this is the case, one must specifically account for parameter estimation error. There are a number of ways to do so. The obvious method is simply to use (14.23)

below when constructing tests, using the obvious sample analogues. That is, if $\hat{F}$, $\hat{B}$, $\hat{S}_{fh}$, $\hat{S}_{hh}$, $\hat{\lambda}_{fh}$, and $\hat{\lambda}_{hh}$ are consistent estimates of their population level counterparts (see below), construct test statistics of the form

$$\hat{\Omega}^{-1/2}P^{-1/2}\Sigma_{t=R}^{R+P-1}(f_{t+1}(\hat{\beta}_t) - \theta), \text{ where} \tag{14.23}$$

$$\hat{\Omega} = \hat{S}_{ff} + \hat{\lambda}_{fh}(\hat{F}\hat{B}\hat{S}'_{fh} + \hat{S}_{fh}\hat{B}'\hat{F}') + \hat{\lambda}_{hh}(\hat{F}\hat{B}\hat{S}_{hh}\hat{B}'\hat{F}').$$

In this way, the test statistic is limiting standard normal, and standard normal tables can be used to construct an asymptotically valid test of the null.

To compute the terms in (14.23), begin with the scalars λ_{fh} and λ_{hh}. Since these are both continuous functions of $\pi = \lim P/R$, they can be computed by substituting P/R for π. Since the term B varies with the methodology used to estimate the parameters (that is, OLS, maximum likelihood, GMM, etc.) so will its estimator. When OLS is used to estimate the parameters, as in the zero-mean prediction error example of (14.7), $B = (EX_tX'_t)^{-1}$. Here, a consistent estimate of B could take the form $\hat{B} = (P^{-1}\Sigma_{t=R}^{R+P-1}X_tX'_t)^{-1}$. Similarly, since the term F varies with the function of interest, $f(.)$, so will its estimator. In the zero-mean prediction error example of (14.7), $F = -EX'_t$. A consistent estimate of F could take the form $\hat{F} = -P^{-1}\Sigma_{t=R}^{R+P-1}X'_t$.

The remaining terms, S_{ff}, S_{fh}, and S_{hh}, require a bit more knowledge of the data that is being used. At times, consistent estimates are quite simple. For example, under the conditions assumed for the linear model in (14.6) and the test of zero-mean prediction error in (14.7), $S_{ff} = \sigma^2$, $S_{fh} = -\sigma^2 EX'_t$, and $S_{hh} = \sigma^2 EX_tX'_t$. Consistent estimates of S_{ff}, S_{fh}, and S_{hh} include $\hat{S}_{ff}^0 = P^{-1}\Sigma_{t=R}^{R+P-1}\hat{u}_{t+1}^2$, $\hat{S}_{fh}^0 = -(P^{-1}\Sigma_{t=R}^{R+P-1}\hat{u}_{t+1}^2)(P^{-1}\Sigma_{t=R}^{R+P-1}X'_t)$, and $\hat{S}_{hh}^0 = (P^{-1}\Sigma_{t=R}^{R+P-1}\hat{u}_{t+1}^2)(P^{-1}\Sigma_{t=R}^{R+P-1}X_tX'_t)$. These estimates, however, are based upon knowledge that the disturbances are i.i.d. and conditionally homoskedastic. If the disturbances are i.i.d. but conditionally heteroskedastic, then $\hat{S}_{ff}^0$ remains consistent for S_{ff} but $\hat{S}_{fh}^0$ and $\hat{S}_{hh}^0$ are no longer consistent for S_{fh} and S_{hh}. In this case, consistent estimators include $\hat{S}_{fh}^1 = -P^{-1}\Sigma_{t=R}^{R+P-1}\hat{u}_{t+1}^2X'_t$ and $\hat{S}_{hh}^1 = P^{-1}\Sigma_{t=R}^{R+P-1}\hat{u}_{t+1}^2X_tX'_t$. More generally, if the disturbances are serially correlated and conditionally heteroskedastic, then standard nonparametric kernel estimators, such as the Bartlett or quadratic spectral, can be used to consistently estimate the long-run covariances S_{ff}, S_{fh}, and S_{hh}. See West and McCracken (1998) and McCracken (2000) for further discussion on estimating these quantities.

Such calculations may be quite involved. Fortunately, in some cases, the formula for Ω simplifies. For example, in the test for zero-mean prediction error above, it can be shown that not only does $-FBS'_{fh} = FBS_{hh}B'F'$, but also $-FBS'_{fh} = FBS_{hh}B'F' = S_{ff}$. Using (14.20) we see that Ω can be rewritten as λS_{ff}, where λ is defined as $1 - 2\lambda_{fh} + \lambda_{hh}$. When this is the case, one need only estimate S_{ff} and λ to form a consistent estimate of Ω. This technique can be particularly useful when the test of interest is for either efficiency or zero-mean prediction error. See West and McCracken (1998) for further examples.

As well, judiciously designed regression-based tests, which introduce seemingly irrelevant variables, can cause the usual least squares standard errors to be the appropriate ones. West and McCracken (1998) demonstrate this, building on

related results for in-sample tests in Pagan and Hall (1983) and Davidson and MacKinnon (1984). Leading examples of tests that can be constructed using augmented regression-based tests include those for serial correlation in the forecast errors and forecast encompassing. For an illustration, consider the encompassing test and assume that the putatively encompassing model is linear with vector of predictors $X_{1,t}$. An adjustment for parameter uncertainty can be made using the augmented regression

$$\hat{u}_{1,t+1} = \alpha_0 + \tilde{\alpha}_2 \hat{y}_{2,t+1} + \alpha'_3 X_{1,t} + \text{error term.} \tag{14.24}$$

(Compare to the usual regression in (14.3).) Under mild conditions, the least squares standard errors associated with the estimate of $\tilde{\alpha}_2$ can be appropriate even when those associated with the estimate of α_2 are not. See West and McCracken (1998) for a discussion on how to choose the augmenting variables.

14.3.4. Monte Carlo evidence

Here we briefly summarize some Monte Carlo evidence on the adequacy of the asymptotic approximations referenced above. In choosing the Monte Carlo studies, we restrict attention to ones that evaluate the finite sample size and size adjusted power properties of tests that use estimated regression models. Thus we do not discuss simulations of out-of-sample tests that do not involve the estimation of parameters, such as those in Diebold and Mariano (1995), Harvey, Leybourne, and Newbold (1997, 1998, 1999), and Clark (1999).

Unfortunately, the number of Monte Carlo studies that evaluate predictive ability using estimated regression models is extremely limited. We cite West (1996, 2001a,b), West and McCracken (1998), McCracken (2000), Corradi, Swanson, and Olivetti (2001), and Clark and McCracken (2001). All of these studies consider one-step-ahead forecast errors and use linear DGPs. Although Corradi, Swanson, and Olivetti (2001) allow the disturbances to be autoregressive, the remaining authors restrict attention to i.i.d. disturbances (usually normal). No doubt equally good finite sample behavior would require larger sample sizes for multi-step forecasts and conditionally heteroskedastic disturbances.

The basic results from these papers are as follows. First, test statistics relying on the asymptotic approximation in (14.18) can work quite well even in sample sizes as small as eight. For example, one set of experiments in West (1996) considers certain tests in which parameter estimation error is asymptotically irrelevant (that is, tests in which $\lambda_{fh}(FBS'_{fh} + S_{fh}B'F') + \lambda_{hh}FBS_{hh}B'F' = 0$ in (14.20)). Of 30 nominal 0.05 tests computed from sample splits in which R ranges from 25 to 100 and P from 25 to 175, all but one had an actual size between 0.03 and 0.08 (the remaining actual size was 0.10). Results in the other papers cited in the previous paragraph are generally comparable. There are, however, occasional exceptions. A dramatic one is the regression-based test of encompassing in (14.24), when the rolling scheme is used. West and McCracken (1998) find that sample sizes greater than a thousand are required for test statistics to be reasonably accurately sized. (Peculiarly, the same regression-based test works well with

small sample sizes when the recursive or fixed schemes are used, for reasons that are not clear to us.)

Second, when the approximation calls for adjusting for parameter estimation error (that is, when $\lambda_{fh}(FBS'_{fh} + S_{fh}B'F') + \lambda_{hh}FBS_{hh}B'F' \neq 0$ in (14.20)), test statistics adjusting for parameter estimation error perform better than those that do not, sometimes dramatically so. West (2001a,b) conducts simulations in which parameter estimation affects the limiting distribution of tests of encompassing. In both papers the simulated tests are conducted with and without adjustment ("Without adjustment" means use only S_{ff} [the first term in (14.20)].). Consider, for example, the simulations in West (2001a). P and R ranged from 8 to 512. For nominal 0.05 tests, sizes of adjusted test statistics ranged from 0.04 to 0.10; the range for unadjusted test statistics was 0.04 to 0.25. (In this particular experiment, it can be shown that failure to adjust will asymptotically result in a size greater than 0.05. In other setups, it is possible that failure to adjust will result in asymptotic sizes less than 0.05.) Consistent with the asymptotic theory and with the intuition of Chong and Hendry (1986), the smaller P/R is, the better is the performance of the unadjusted test statistic. For example, for the unadjusted test statistic, with $P = 32$, nominal 0.05 tests behaved as follows: when $R = 16$ (that is, $P/R = 2$), actual size $= 0.208$; when $R = 256$ (that is, $P/R = 0.125$), actual size $= 0.055$. (The comparable figures for the test statistic that adjusts for parameter estimation error are 0.065 for $R = 16$ and 0.050 for $R = 256$.)

Third, when the approximation calls for adjusting for parameter estimation error, size-adjusted power typically is better for adjusted than for unadjusted test statistics. Both Clark and McCracken (2001) and McCracken (2000) conduct simulations in which parameter estimation affects the limiting distribution. In both papers, the asymptotically valid version of the test has greater size-adjusted power than the asymptotically invalid version. Consider, for example, one of the setups in McCracken (2000), in which $R = 432$ and $P = 87$ (that is, $P/R = 0.20$), the recursive scheme is used, and the nominal size of the test is 0.01. As the deviation from the null decreases across four experiments, the adjusted tests have size-adjusted powers of 1.000, 0.988, 0.748, and 0.118, while the unadjusted tests have size-adjusted powers of 0.999, 0.948, 0.383, and 0.054 respectively.

In addition, for size-adjusted tests, power is increasing in P/R. This tendency occurs in each of the simulations conducted by Corradi, Swanson, and Olivetti (2001), Clark and McCracken (2001), and McCracken (2000). For example, in the first of two simulations conducted by Corradi, Swanson, and Olivetti (2001) they construct tests for equal MSE estimating parameters associated with both stationary regressors and those with a unit root. They find that nominal 0.05 tests have size-adjusted powers that range across 0.170, 0.261, 0.466, 0.632, 0.764, and 0.847 when $R = 50$ and $P = 25, 50, 100, 150, 200, 250$ (that is, $P/R = 0.50, 1, 2, 3, 4, 5$) respectively.

14.4. CONCLUSION

We hope that we have made four contributions in this chapter. The first is simply to introduce the methodology of out-of-sample hypothesis testing. We think this

is important. The applied papers cited in section 14.2 convince us that out-of-sample methods provide an insight into econometric models that is not easily obtained with standard in-sample methods.

The second contribution is to emphasize that economic forecasts and forecast errors usually are generated using estimated regression models. Because of this feature, test statistics constructed using forecasts and forecast errors often have statistical properties that differ from those of statistics constructed when the population level forecasts and forecast errors are known. For those tests that are asymptotically normal, we show that the limiting variance may depend on uncertainty about the parameters used to make forecasts.

Our third contribution is to outline how to feasibly account for the uncertainty introduced when estimated regression parameters are used to construct forecasts. We describe circumstances under which such estimation error washes out asymptotically and thus can be ignored in large samples. For other circumstances, we outline how adjustments can be made.

The fourth, and perhaps most important, contribution that our chapter can make is to promote future use of and research on out-of-sample tests of predictive ability. There are more than a few unanswered questions regarding out-of-sample tests of predictive ability. Perhaps the most obvious is how to choose the sample split parameter π (defined in (14.9)). The Monte Carlo results in section 14.3.4 seem to indicate that this can play a significant role in determining both the size and power properties of the test. Further Monte Carlo results and perhaps even local asymptotics may shed light on the optimal choice of π. Steckel and Vanhonacker (1993) provide some suggestions, but conclude that in finite samples a split where $P/R \approx 1$ is not "too" suboptimal.

Other lines of research include extending the work of Corradi, Swanson, and Olivetti (2001) in ways that allow for a broader range of tests than just for equal forecast accuracy between two nonnested models. Examples include tests of efficiency, encompassing, serial correlation in the forecast errors and zero-mean prediction error when cointegrating relationships exist within the estimated regression model(s). Moreover, since tests of causality often involve regression models that include cointegrating relationships (Stock and Watson, 1988) it would be useful to extend their work along the lines of that by McCracken (1999) and Clark and McCracken (2001). In that way, tests of equal forecast accuracy and encompassing between two nested models can be constructed when cointegrating relationships exist.

Another line of work involves tests of predictive ability when nonparametric methods are used to estimate regression functions. Swanson and White (1997a, b), Chung and Zhou (1996), and Jaditz and Sayers (1998) each construct tests of predictive ability based upon forecasts and forecast errors estimated using series-based, kernel-based, and local-linear nonparametric methods respectively. Since each of these methods use regression estimators that have slower rates of convergence than parametric estimates it seems likely that they, too, would affect the distributions of out-of-sample tests of predictive ability.

Other potential questions include determining the importance of parameter uncertainty in the evaluation of probability or interval forecasts. There is also the

question of whether out-of-sample methods should be used at all. Fair and Shiller (1990) suggest that this methodology provides a protection against data-mining and overfitting. It would be useful to determine whether their heuristic arguments in favor of out-of-sample methods can be validated analytically. Another question is whether standard information criteria can be used as a consistent testing methodology in an out-of-sample environment. In particular, information criteria are sometimes used to choose among multiple nested and nonnested models based upon out-of-sample forecast errors. Although Swanson (1998) argues in its favor, there is no theoretical evidence to support its use.

Acknowledgments

References

Alpert, A.T. and J.B. Guerard (1988). Employment, unemployment and the minimum wage: A causality model. *Applied Economics*, 20, 1453–64.

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, M.J., A.P.L. Minford, and J. Riley (1996). On comparing macroeconomic models using forecast encompassing tests. *Oxford Bulletin of Economics and Statistics*, 58, 279–305.

Ashley, R. (1981). Inflation and the distribution of price changes across markets: a causal analysis. *Economic Inquiry*, 19, 650–60.

Ashley, R. (1998). A new technique for postsample model selection and validation. *Journal of Economic Dynamics and Control*, 22, 647–65.

Ashley, R., C.W.J. Granger, and R. Schmalensee (1980). Advertising and aggregate consumption: an analysis of causality. *Econometrica*, 48, 1149–67.

Bates, J.M. and C.W.J. Granger (1969). The combination of forecasts. *Operational Research Quarterly*, 20, 451–68.

Berger, A. and S. Krane (1985). The informational efficiency of econometric model forecasts. *Review of Economics and Statistics*, 67, 128–34.

Blomberg, S.B. and G.D. Hess (1997). Politics and exchange rate forecasts. *Journal of International Economics*, 43, 189–205.

Bradshaw, G.W. and D. Orden (1990). Granger causality from the exchange rate to agricultural prices and export sales. *Western Journal of Agricultural Economics*, 15, 100–10.

Bram, J. and S. Ludvigson (1998). Does consumer confidence forecast household expenditure? A sentiment horse race. *Federal Reserve Bank of New York Economic Policy Review*, 4, 59–78.

Brandt, J.A. and D. Bessler (1983). Price forecasting and evaluation: an application in agriculture. *Journal of Forecasting*, 2, 237–48.

Breen, W., L.R. Glosten, and R. Jagannathan (1989). Economic significance of predictable variations in stock index returns. *Journal of Finance*, 44, 1177–89.

Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 75, 1–3.

Chen, X. and N.R. Swanson (1996). Semiparametric ARX neural network models with an application to forecasting inflation. Manuscript, London School of Economics and Purdue University.

Chinn, M.D. and R.A. Meese (1995). Banking on currency forecasts: How predictable is change in money? *Journal of International Economics*, 38, 161–78.

Chong, Y.Y. and D.F. Hendry (1986). Econometric evaluation of linear macro-economic models. *Review of Economic Studies*, 53, 671–90.

Christ, C. (1956). Aggregate econometric models. *American Economic Review*, 66, 385–408.

Chung, Y.P. and Z.G. Zhou (1996). The predictability of stock returns – a nonparametric approach. *Econometric Review*, 15, 299–330.

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 (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105, 85–110.

Corradi, V., N.R. Swanson and C. Olivetti (2001). Predictive ability with cointegrated variables. *Journal of Econometrics*, 105, 315–58.

Cumby, R.E. and D.M. Modest (1986). Testing for market timing ability: A framework for forecast evaluation. *Journal of Financial Economics*, 19, 169–89.

Davidson, R. and J.G. MacKinnon (1984). Model specification tests based on artificial linear regressions. *International Economic Review*, 25, 241–62.

Dawid, A.P. (1984). Present position and potential developments: some personal views. Statistical theory. The prequential approach. *Journal of the Royal Statistical Society, Series A*, 147, 278–92.

Day, T.E. and C.M. Lewis (1992). Stock market volatility and the information content of stock index options. *Journal of Econometrics*, 52, 267–87.

Diebold, F.X. (1989). Forecast combination and encompassing: reconciling two divergent literatures. *International Journal of Forecasting*, 5, 589–92.

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, 603–10.

Durbin, J. (1970). Testing for serial correlation in least squares regression when some of the regressors are lagged dependent variables. *Econometrica*, 38, 410–21.

Engel, C. (1994). Can the Markov switching model forecast exchange rates? *Journal of International Economics*, 36, 151–65.

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.

Ericsson, N.R. and J. Marquez (1993). Encompassing the forecasts of U.S. trade balance models. *Review of Economics and Statistics*, 75, 19–31.

Fair, R.C. (1980). Estimating the predictive accuracy of econometric models. *International Economic Review*, 21, 355–78.

Fair, R.C. (1999). Evaluating the information content and money making ability of forecasts from exchange rate equations. Manuscript, Yale University.

Fair, R.C. and R. Shiller (1989). The informational content of ex ante forecasts. *Review of Economics and Statistics*, 71, 325–31.

Fair, R.C. and R. Shiller (1990). Comparing information in forecasts from econometric models. *American Economic Review*, 80, 375–89.

Goldberger, A.S. (1959). *Impact Multipliers and Dynamic Properties of the Klein–Goldberger Model*. Amsterdam: North-Holland.

Goldberger, A.S. (1991). *A Course in Econometrics*. Cambridge: Harvard University Press.

Granger, C.W.J. and M. Deutsch (1992). Comments on the evaluation of policy models. *Journal of Policy Modeling*, 14, 497–516.

Granger, C.W.J. and P. Newbold (1977). *Forecasting Economic Time Series*. London: Academic Press.

Granger, C.W.J. and M.H. Pesaran (2000). A decision theoretic approach to forecast evaluation. In W.S. Chan, W.K. Lin, and H. Tong (eds.), *Statistics and Finance: An Interface*. London: Imperial College Press, 261–78.

Granger, C.W.J. and R. Ramanathan (1984). Improved methods of combining forecasts. *Journal of Forecasting*, 3, 197–204.

Granger, C.W.J. and N.R. Swanson (1997). An introduction to stochastic unit-root processes. *Journal of Econometrics*, 80, 35–62.

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.

Harvey, D.I., S.J. Leybourne, and P. Newbold (1999). Forecast evaluation tests in the presence of ARCH. *Journal of Forecasting*, 18, 435–45.

Hatanaka, M. (1974). A simple suggestion to improve the Mincer–Zarnowitz criterion for the evaluation of forecasts. *Annals of Economic and Social Measurement*, 3, 521–4.

Henriksson, R.D. and R.C. Merton (1981). On market timing and investment performance II: Statistical procedures for evaluating forecasting skills. *Journal of Business*, 54, 513–33.

Hoffman, D. and A. Pagan (1989). Practitioners corner: post-sample prediction tests for generalized method of moments estimators. *Oxford Bulletin of Economics and Statistics*, 51, 333–43.

Jaditz, T. and C.L. Sayers (1998). Out-of-sample forecast performance as a test for nonlinearity in time series. *Journal of Business and Economic Statistics*, 16, 110–17.

Keane, M.P. and D.E. Runkle (1989). Are economic forecasts rational? *Federal Reserve Bank of Minneapolis Quarterly Review*, 13.

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

Kitamura, Y. (1999). Predictive inference and the bootstrap. Manuscript, University of Wisconsin.

Klein, L.R. (1992). The test of a model is its ability to predict. Manuscript, University of Pennsylvania.

Kuan, C. and T. Liu (1995). Forecasting exchange rates using feedforward and recurrent neural networks. *Journal of Applied Econometrics*, 10, 347–64.

Leitch, G. and J.E. Tanner (1991). Economic forecast evaluation: profits versus the conventional error measures. *American Economic Review*, 81, 580–90.

Lo, A.W. and A.C. MacKinlay (1990). Data snooping biases in tests of financial asset pricing models. *Review of Financial Studies*, 3, 431–67.

Lo, A.W. and A.C. MacKinlay (1997). Maximizing predictability in the stock and bond markets. *Macroeconomic Dynamics*, 1, 118–58.

Lopez, J.A. (2001). Evaluating the predictive accuracy of volatility models. *Journal of Forecasting*, 20, 87–109.

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

McCracken, M.W. (1999). Asymptotics for out-of-sample tests of causality. Manuscript, University of Missouri-Columbia.

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

McCulloch, R. and P.E. Rossi (1990). Posterior, predictive, and utility-based approaches to testing the arbitrage pricing theory. *Journal of Financial Economics*, 28, 7–38.

McNees, S.K. (1978). The rationality of economic forecasts. *American Economic Review; Papers and Proceedings of the Ninetieth Annual Meeting of the American Economic Association*, 68, 301–5.

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, 933–48.

Merton, R.C. (1981). On market timing and investment performance. I. An equilibrium theory of value for market forecasts. *Journal of Business*, 54, 363–406.

Mincer, J. and V. Zarnowitz (1969). The evaluation of economic forecasts. In J. Mincer (ed.), *Economic Forecasts and Expectations*. New York: National Bureau of Economic Research, 3–46.

Morgan, W.A. (1939). A test for significance of the difference between two variances in a sample from a normal bivariate population. *Biometrika*, 31, 13–19.

Murphy, A.H. and H. Daan (1985). Forecast evaluation. In A.H. Murphy and R.W. Katz (eds.), *Probability, Statistics and Decision Making in the Atmospheric Sciences*. Boulder: Westview Press, 379–437.

Oliner, S., G. Rudebusch, and D. Sichel (1995). New and old models of business investment: A comparison of forecasting performance. *Journal of Money, Credit and Banking*, 27, 806–26.

Pagan, A.R. and A.D. Hall (1983). Diagnostic tests as residual analysis. *Econometric Review*, 2, 159–218.

Pagan, A.R. and G.W. Schwert (1990). Alternative models for conditional stock volatility. *Journal of Econometrics*, 45, 267–90.

Park, T. (1990). Forecast evaluation for multivariate time-series models: the U.S. cattle market. *Western Journal of Agricultural Economics*, 15, 133–43.

Pesaran, M.H. and A. Timmermann (1992). A simple nonparametric test of predictive performance. *Journal of Business and Economic Statistics*, 10, 561–5.

Pesaran, M.H. and A. Timmermann (1994). A generalization of the nonparametric Henriksson–Merton test of market timing. *Economics Letters*, 4, 1–7.

Pesaran, M.H. and A. Timmermann (1995). Predictability of stock returns: robustness and economic significance. *Journal of Finance*, 50, 1201–28.

Randles, R. (1982). On the asymptotic normality of statistics with estimated parameters. *Annals of Statistics*, 10, 462–474.

Steckel, J. and W. Vanhonacker (1993). Cross-validating regression models in market research. *Marketing Science*, 12, 415–27.

Stock, J.H. and M.W. Watson (1988). Interpreting the evidence on money-income causality. *Journal of Econometrics*, 40, 161–82.

Stock, J.H. and M.W. Watson (1993). A procedure for predicting recessions with leading indicators: Econometric issues and recent experience. In J.H. Stock and Mark W. Watson (eds.), *Business Cycles, Indicators and Forecasting*. Chicago: University of Chicago Press, 95–153.

Stock, J.H. and M.W. Watson (1999a). Forecasting inflation. *Journal of Monetary Economics*, 44, 293–335.

Stock, J.H. and M.W. Watson (1999b). Diffusion indices. *NBER Working Paper* #6702.

Swanson, N.R. (1998). Money and output viewed through a rolling window. *Journal of Monetary Economics*, 41, 455–73.

Swanson, N.R. and H. White (1995). A model-selection approach to assessing the information in the term structure using linear models and artificial neural networks. *Journal of Business and Economic Statistics*, 13, 265–75.

Swanson, N.R. and H. White (1997a). A model-selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks. *Review of Economics and Statistics*, 79, 265–75.

Swanson, N.R. and H. White (1997b). Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models. *International Journal of Forecasting*, 13, 439–61.

Tegene, A. and F. Kuchler (1994). Evaluating forecasting models of farmland prices. *International Journal of Forecasting*, 10, 65–80.

West, K.D. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64, 1067–84.

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

West, K.D. (2001b). Encompassing tests when no model is encompassing. *Journal of Econometrics*, 105, 287–308.

West, K.D. and D. Cho (1995). The predictive ability of several models of exchange rate volatility. *Journal of Econometrics*, 69, 367–91.

West, K.D., H.J. Edison, and D. Cho (1993). A utility-based comparison of some models of exchange rate volatility. *Journal of International Economics*, 35, 23–45.

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, 1097–126.