Tests for Forecast Encompassing When Forecasts Depend on Estimated Regression Parameters

Kenneth D. West
Department of Economics, University of Wisconsin, Madison, WI 53706 (*kdwest@facstaff.wisc.edu*)

This article presents analytical and simulation results on the properties of two tests for forecast encompassing, allowing throughout for dependence of the forecasts on estimated regression parameters. One test, which was intended for forecasts that do not depend on regression parameters, was developed by Harvey, Leybourne, and Newbold. This test works relatively well when the size of the sample of forecast errors is very small. A second test, which explicitly accounts for uncertainty about the regression parameters, otherwise is comparable or preferable.

KEY WORDS: Generated regressors; Hypothesis testing; Nonnested models; Out-of-sample; Prediction.

In a recent article, Harvey, Leybourne, and Newbold (1998, henceforth HLN) developed and evaluated tests for forecast encompassing. The theory and simulations of HLN assume that forecast errors are observed without error. But many if not most economic forecasts rely on regression estimates. Forecast errors thus often are regression residuals that are contaminated by error in estimation of the regression parameters.

This note considers forecast encompassing when forecasts rely on regression estimates. It concludes that even when forecasts are based on regression estimates, the test statistic proposed by HLN (1998) is relatively attractive under the following condition: The size of the sample of forecast errors is very small, as measured either by number of observations (say, $n \leq 8$) or as a fraction of the size of the sample used to compute the regression estimates (say, n less than a tenth of the regression sample size). But otherwise, attempting to apply this statistic when forecasts rely on regression estimates—an application not considered or recommended by HLN (1998), I hasten to point out—probably is inadvisable. A test statistic that explicitly accounts for the dependence of the forecast errors on regression estimates seems comparable or preferable.

For expositional clarity, I use an example with univariate least squares forecasting models, one-step-ahead forecast errors, and forecasts made using a single estimate of the regression parameters. The appendix presents the generalization to nonlinear and multivariate models, multistep forecasts, and forecasting schemes in which parameter estimates are updated as new data become available so that different forecasts use different regression estimates.

1. ASYMPTOTIC RESULTS

In the example I use, the investigator is comparing two models for a scalar variable y_t. The two models are

$$y_t = X_{1t}\beta_1 + e_{1t}$$
$$y_t = X_{2t}\beta_2 + e_{2t}. \tag{1}$$

In (1), all variables are scalars, and β_1 and β_2 are unknown parameters. The variables are assumed to be stationary and well behaved in the sense made precise by Diebold and Mariano (1995), HLN (1998), and West and McCracken (1998). In this simple example I also impose the conditions that the vector of regression disturbances $(e_{1t}, e_{2t})'$ is independently and identically distributed, with each disturbance uncorrelated with the corresponding regressor (i.e., $Ee_{1t}X_{1t} = Ee_{2t}X_{2t} = 0$).

To compare the two models, one performs an encompassing test. Let

$$d_t = e_{1t}^2 - e_{1t}e_{2t}. \tag{2}$$

As explained by HLN (1998), if Model (1) encompasses Model (2), $Ed_t = 0$, and that is the null.

If β_1 and β_2 were observed, one could test H_0: $Ed_t = 0$ by examining

$$\tilde{d} \equiv n^{-1} \sum_{t=R+1}^{R+n} d_t \equiv n^{-1} \sum_{t=R+1}^{R+n} (e_{1t}^2 - e_{1t}e_{2t}). \tag{3}$$

(See the following paragraphs for why the sample starts at $R+1$.) Let S denote the variance of d_t. HLN (1998) noted that it follows from Diebold and Mariano (1995) that, under standard assumptions about forecast errors,

$$\sqrt{n}(\tilde{d} - Ed_t) \sim_A \mathrm{N}(0, S), S \equiv E(d_t - Ed_t)^2. \tag{4}$$

This suggests obtaining a standard error as the square root of the sample variance of d_t. The simulations of HLN (1998) show, however, that using degrees-of-freedom adjustment and using critical values from a t rather than a normal distribution result in distinctive improvement in test statistics.

What are the implications of β_1 and β_2 being unknown? Evidently, they will be estimated before forecasts are made. So suppose that data from $t = 1, \ldots, R$ are used to obtain least squares estimates $\hat{\beta}_1$ and $\hat{\beta}_2$. These estimates are then used to

make forecasts from $t = R+1, \ldots, R+n$. (I take as given the split of the sample into R and n.) One examines not $\tilde{d}$ but

$$\bar{d} \equiv n^{-1} \sum_{t=R+1}^{R+n} (\hat{e}_{1t}^2 - \hat{e}_{1t}\hat{e}_{2t})$$
$$\hat{e}_{1t} = y_t - X_{1t}\hat{\beta}_1,\ \hat{e}_{2t} = y_t - X_{2t}\hat{\beta}_2. \quad (5)$$

Define

$$V_\beta = (2 \times 2) \text{ asymptotic variance–covariance matrix of the estimator of } \beta \equiv (\beta_1, \beta_2)'$$
$$D = E\frac{\partial d_t}{\partial \beta'} = (1 \times 2) \text{ expectation of the vector of derivatives of } d_t \text{ with respect to } \beta'$$
$$D = (-2Ee_{1t}X_{1t} + Ee_{2t}X_{1t}, Ee_{1t}X_{2t})$$
$$= (Ee_{2t}X_{1t}, Ee_{1t}X_{2t}),$$
$$\pi = \text{limit as sample size goes to infinity of } \frac{n}{R}, \pi < \infty. \quad (6)$$

Under the conditions of West and McCracken (1998),

$$\sqrt{n}(\bar{d} - Ed_t) \sim_A \mathrm{N}(0, \Omega), \Omega = S + \pi DV_\beta D'$$
$$\equiv E(d_t - Ed_t)^2 + \pi DV_\beta D'. \quad (7)$$

On comparing (7) and (4), we see that parameter estimation introduces extra uncertainty because $\pi DV_\beta D'$ is nonnegative and except in very special cases will be strictly positive. Evidently, using S (or a consistent estimate of S) to perform inference, which is appropriate when there is no parameter estimation error, will result in too many asymptotic rejections at any specified significance level. In considering use of HLN's (1998) test, the question is whether the distortion is trivial and so can be ignored in practice. I first consider the asymptotic formulas (4) and (7) analytically, and then turn to simulations to quantify the effects of parameter uncertainty in finite samples.

One simple condition that ensures that the distortion is small is that π is small because, for arbitrarily small π, $\pi DV_\beta D'$ is arbitrarily small compared to S. When π is small, one keeps $n \ll R$ as the sample size grows so that many more observations are used to obtain the estimates of β than are used to obtain the estimate of Ed_t. As one might expect, in such a case uncertainty about β will be small compared to uncertainty that would be present even if β were known. If one uses n/R as the obvious finite-sample analogue to π, then a range of values of π is suggested, some small [e.g., $n/R \approx .15$ as in Ericcson and Marquez (1993)], some moderate [e.g., $n/R \approx .4$ as in Cooper (1972)], some, especially in financial applications, large [e.g., the range of values of n/R is from about 5 to 18 as in Engle, Hong, and Kane (1990)].

Of course, the term $\pi DV_\beta D'$ is small for any π if V_β is small: Ceteris paribus, the smaller the variance of the estimator of β, the less important is such variance for inference about Ed_t. [Note, however, that (in contrast to π) one is typically not free to vary V_β without also varying D and $E(d_t - Ed_t)^2$.] To gauge the magnitude of $DV_\beta D'$, let us impose the null hypothesis that Model (1) encompasses Model (2). Then e_{1t} cannot be predicted by Model (2), so

$$D = (Ee_{2t}X_{1t}\ 0) \Rightarrow DV_\beta D'$$
$$= (Ee_{2t}X_{1t})^2 V_\beta(1,1)$$
$$= (Ee_{2t}X_{1t})^2 (EX_{1t}^2)^{-2} (EX_{1t}^2 e_{1t}^2). \quad (8)$$

[In (8), the usual heteroscedasticity-robust formula has been used for the asymptotic variance of the least squares estimator of β_1.] To get a precise, and simple, statement about the relative size of the two terms in (7), let us further assume that $\beta_2 = 0$ ($\Rightarrow e_{2t} = X_{1t}\beta_1 + e_{1t}$) and that X_{1t} and e_{1t} are iid normal. With some algebra, one can then establish that

$$DV_\beta D' = S \Rightarrow \sqrt{n}(\bar{d} - Ed_t) \sim_A \mathrm{N}(0, \Omega)$$
$$\Omega = (1+\pi)S \equiv (1+\pi)E(d_t - Ed_t)^2$$
$$= (1+\pi)Ed_t^2. \quad (9)$$

The final equality holds since $Ed_t = 0$ under the null.

It is easy to calibrate the asymptotic magnitude of the distortion from using S (or a consistent estimate of S) to perform inference. If (say) $\pi = .1$, asymptotic t statistics using $\sqrt{S}$ rather than $\sqrt{(1+\pi)S}$ will be too small by a factor of $\sqrt{1.1} \approx 1.03$, implying that nominal 5% tests will have actual size of about 6%. This is arguably a small distortion; it affirms Chong and Hendry's (1986) view that, if n/R is small, one can sometimes safely abstract from error in estimation of parameters when performing encompassing tests. But for larger values of π, the distortion is larger. If $\pi =$ (say).5, nominal 5% t tests using $\sqrt{S}$ rather than $\sqrt{(1+\pi)S}$ will asymptotically have actual size of about 11. If $\pi = 2$, the implied size is about 26. As noted previously, many applications do involve ratios of prediction to regression samples that are moderate or large, indicating that accounting for parameter estimation error will be important in samples large enough for the asymptotic approximation to be accurate.

The simple result (9) requires assumptions that will be difficult to defend in most applications—for example, that X_{1t} is iid. As discussed in the appendix, the asymptotic result (7) holds quite generally. It may help to note here that under the null that Model (1) encompasses Model (2), (7) typically takes a relatively simple form. Suppose that Model (1) relies on a parameter vector $\beta_1 (k_1 \times 1)$ Model (2) a parameter vector β_2 (the dimension is not relevant) with corresponding prediction errors e_{1t} and e_{2t}. Let $\partial e_{1t}/\partial \beta_1$ be the $(k_1 \times 1)$ derivative of e_{1t} with respect to β_1. For example, in the linear model $y_t = X_{1t}'\beta_1 + e_{1t}$, $\partial e_{1t}/\partial \beta_1 = -X_{1t}$. Let $V_{\beta 1}$ denote the $k_1 \times k_1$ asymptotic variance–covariance matrix of β_1. Then, under the null, the asymptotic variance of d_t is

$$Ed_t^2 + \pi[Ee_{2t}(\partial e_{1t}/\partial \beta_1)]' V_{\beta 1} [Ee_{2t}(\partial e_{1t}/\partial \beta_1)].$$

The asymptotic variance is more complicated if the prediction horizon is multiperiod, or if one uses what the appendix calls "recursive" or "rolling" schemes for forecasting. See the appendix for details.

2. SIMULATION RESULTS

To get a sense of whether the asymptotic approximation is helpful in samples of typical size, I performed a small Monte Carlo experiment. As in HLN (1998), I tried prediction sample sizes $n = 8, 16, 32, 64, 128$, and 256. For each choice of n, I tried a range of values of R, subject to the constraint that $n + R \leq 768$: $n/R = 2, 1, .5, .25, .125, .0625$. (The upper bound of 768 was imposed to limit the amount of computation. For $n = 256$, for example, this choice allows $n/R = 2$, 1, and .5, but not $n/R = .25, .125$, or .0625.) The range of values of n/R tends toward the smaller side of values found in the empirical work that I am familiar with (see the preceding references). I lean toward smaller values because the asymptotic theory suggests that there will be a smaller distortion from ignoring dependence on estimated regression parameters for smaller values of n/R. The number of replications was 10,000. I generated the data according to (1), with $\beta_2 = 0$, $\beta_1 = 1$, and $(e_{1t}, X_{1t}, X_{2t})'$ iid normal with diagonal variance–covariance matrix with diagonal elements (1, 1, 2).

Using data from 1 to R, least squares was used to obtain estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ and $\hat{V}_\beta(1, 1)$, the usual heteroscedasticity-consistent estimate of $V_\beta(1, 1)$. (No heteroscedasticity is present in the data, but to be conservative I assume that the investigator does not know this.) I used sample averages of $\hat{e}_{2t}X_{1t}$ and of $(\hat{d}_t - \bar{d})^2$ to construct $\hat{D}(1)$ and $\hat{S}$, where the sample was over the n forecast errors. Two t tests were performed. The first mimicked HLN (1998), making a degrees-of-freedom adjustment to $\hat{S}$ and using critical values from a $t(n-1)$ distribution. The second used conventional asymptotics:

$$\text{Test 1: Compute } \frac{\sqrt{n}\,\bar{d}}{\sqrt{\frac{n}{n-1}\hat{S}}};$$

$$\text{compare to } t(n-1) \text{ distribution.} \qquad (10a)$$

$$\text{Test 2: Compute } \frac{\sqrt{n}\,\bar{d}}{\sqrt{\hat{\Omega}}};$$

$$\text{compare to N(0, 1) distribution,}$$

$$\hat{\Omega} = \hat{S} + \left(\frac{n}{R}\right)\hat{D}(1)^2\hat{V}_\beta(1, 1). \qquad (10b)$$

Results for two-sided nominal 5% tests are presented in Table 1. (As documented in an additional appendix that is available on request, results were similar for 10% tests; for 5 and 10% tests that use $t(6)$ rather than normally distributed data, for tests that construct $\hat{\Omega}$ imposing homoscedasticity in estimation of $V_\beta(1, 1)$, and for tests that do not impose $Ee_{1t}X_{2t} = 0$ in estimation of $DV_\beta D'$.) The "8.3" near the upper left of the table, for example, indicates that in about 830 of the 10,000 simulations, Equation (10a) was greater than 2.365 in absolute value [2.365 is the critical value for a two-sided .05 test for a $t(7)$ distribution].

As may be seen from the upper left corner of the table, the statistic (10a), which does not explicitly account for sampling error in estimation of β_1 and β_2, is more accurate when $n = 8$ and $n/R = 1, 2$. Neither statistic is especially accurate when $n = 8$ and $n/R = 2$, a perhaps unsurprising finding since the size of the sample used to estimate the β's is so small ($R = 4$).

Table 1. Empirical Sizes of Nominal 5%-Level Tests for Forecast Encompassing

n	Test statistic	$\frac{n}{R}$ = 2	1	.5	.25	.125	.0625
8	Eq. (10a)	8.3	5.9	5.0	4.3	3.7	3.6
	Eq. (10b)	9.8	6.4	5.4	5.3	6.3	7.7
16	Eq. (10a)	15.5	10.8	7.8	5.9	5.1	4.8
	Eq. (10b)	8.0	5.1	4.1	4.5	5.3	6.1
32	Eq. (10a)	20.8	14.2	9.6	6.5	5.5	5.1
	Eq. (10b)	6.5	5.0	4.2	4.3	5.0	5.5
64	Eq. (10a)	23.5	15.5	10.3	7.9	6.6	
	Eq. (10b)	5.5	4.7	4.8	5.2	5.4	
128	Eq. (10a)	24.7	15.7	10.6	8.1		
	Eq. (10b)	5.3	4.7	4.8	5.2		
256	Eq. (10a)	25.2	16.7	10.8			
	Eq. (10b)	5.0	4.7	4.8			

NOTE: 1. This table presents empirical sizes of nominal .05 two-sided tests for H_0: $Ed_t = 0$, $d_t \equiv E(e_{1t}^2 - e_{1t}e_{2t})$. The model is given in (1), with $\beta_1 = 1$, $\beta_2 = 0$, $(e_{1t}, X_{1t}, X_{2t})'$ iid normal with variance–covariance matrix = diag(1,1,2). 2. "n" is the number of one-step-ahead forecast errors used to compute $\bar{d}$; $\bar{d}$ is the sample average of $\hat{e}_{1t}^2 - \hat{e}_{1t}\hat{e}_{2t}$, where $\hat{e}_{it}$ is the forecast error when the least squares estimate of β_i is used for forecasting [see (5)]; R is the sample size used to compute the least squares estimates of β_1 and β_2. Thus in the row labeled "8," the values of R corresponding to the six values of n/R are 4, 8, 16, 32, 64, and 128. 3. Equation (10a) is a statistic that would be asymptotically valid as $n \to \infty$ if the e_{it}'s were observed, rather than computed as regression residuals. Because regression residuals are used and $0 < \pi \equiv \lim_{R,n\to\infty} n/R$, Equation (10a) will asymptotically have an actual size greater than .05. Equation (10b) is a statistic that is asymptotically valid as $R, n \to \infty$ and takes account of the fact that $\bar{d}$ is constructed from regression residuals. 4. Equation (10a) uses critical values from a $t(n-1)$ distribution, Equation (10b) from an N(0, 1) distribution. The entries "8.3" and "9.8" in row $n = 8$, column $n/R = 2$, for example, indicates that in about 830 of the 10,000 samples, the Equation (10a) test statistic was greater than 2.365 in absolute value [2.365 is the critical value for a $t(7)$ distribution]; in about 980 of the 10,000 samples, the Equation (10b) test statistic was greater than 1.96 in absolute value.

The statistic (10b), which explicitly accounts for such sampling error, typically is more accurate when $n \geq 16$ and $n/R \geq .125$, though there are exceptions ($n = 16$, $n/R = .125$) and a number of cases in which the ranking depends on the loss function for over- versus under-rejection. In connection with the loss function, my own sense is that most economists would prefer to underreject slightly rather than overreject sharply, and that sense underlies my interpretation of (10b) as preferable (e.g., for $n = 256$, nominal sizes of 4.7 and 4.8 are preferable to 16.7 and 10.8). By the same token, the last column of the table suggests that, for $n/R = .0625$, the test (10a) is, overall, preferable. This is consistent with the previous interpretation that abstracting from error in estimation of regression parameters is safe when the ratio of forecast size to regression size is small.

For larger n/R, the asymptotic tendency of Equation (10a) to reject too much is reflected in the simulation. Indeed, for $n \geq 64$, the sizes of the statistic (10a) predicted by the asymptotic approximation are roughly reflected. For example, the asymptotic size is about 26 for $\lim_{R,n\to\infty} n/R \equiv \pi = 2$, versus the 23.5–25.2 range reported in the table.

3. SUMMARY

Many if not most economic forecasts are based on regression estimates. In performing an encompassing test using such forecasts, the statistic proposed by HLN (1998) is relatively attractive for very small n (say, $n \leq 8$) or when n is very

small compared to the size of the sample used to compute the regression estimates (say, $n/R < .1$). But otherwise, asymptotic and simulation results argue for applying instead a test statistic that explicitly accounts for the dependence of the forecast errors on regression estimates. Unfortunately, perceptible size distortions remain in samples of relevant size, and alternative procedures that lead to more accurately sized tests are highly desirable.

ACKNOWLEDGMENTS

I thank Stanislav Anatolyev, Yuichi Kitamura, Paul Newbold, Ka-fu Wong, two anonymous referees, and the editor (Jeffrey Wooldridge) for helpful comments and discussions, Stanislav Anatolyev for excellent research assistance, and the National Science Foundation for financial support.

APPENDIX: ASYMPTOTIC RESULTS

This appendix sketches the general version of the asymptotic distribution of the estimator of Ed_t, taking account of dependence of the prediction errors on estimated parameters. Technical conditions and further discussion may be found in West (1996) and West and McCracken (1998).

Let y_t denote the scalar variable being forecast by two models. As previously, write the parameter vector from model i as β_i and stack the two vectors in a $(k \times 1)$ vector $\beta \equiv (\beta_1', \beta_2')'$. Let the total amount of data available be $R+n+h-1$, where "h" is the forecast horizon ($h = 1$ in the preceding analysis). Three schemes for obtaining parameter estimates figure prominently in the literature. The schemes must be distinguished because asymptotic results vary across the three. The *fixed* scheme was the one illustrated previously. Data from 1 to R are used to estimate β_1 and β_2 and this single set of estimates is used in all predictions. In the *rolling* scheme, the investigator rolls through the sample, first using data from 1 to R to estimate the β_i's and predict y_{R+h}, then data from 2 to $R+1$ are used to estimate the β_i's and this new set of estimates is used to predict $y_{R+1+h}, \ldots$, and finally data from n to $R+n-1$. In the *recursive* scheme, the sample size used to estimate the β_i's grows: The investigator first uses from 1 to R to estimate the β_i's and predict y_{R+h}, then estimates using data from 1 to $R+1$ and uses the new estimates to predict $y_{R+1+h}, \ldots$, and finally estimates using data from 1 to $R+n-1$ to predict $y_{R+n+h-1}$.

As previously, define $\pi < \infty$ as $\pi = \lim_{R,n\to\infty} n/R$. Define the scalars λ_{dh} and λ_{hh} as follows:

Sampling scheme	λ_{dh}	λ_{hh}
recursive	$1-\pi^{-1}\ln(1+\pi)$	$2[1-\pi^{-1}\ln(1+\pi)]$
rolling, $\pi \le 1$	$\frac{\pi}{2}$	$\pi - \frac{\pi^2}{3}$
rolling, $\pi > 1$	$1-\frac{1}{2\pi}$	$1-\frac{1}{3\pi}$
fixed	0	π.

(A.1)

I assume predictions at time t rely on an estimate of $\beta \equiv (\beta_1', \beta_2')$ that can be written $\hat{\beta}(t) = B(t)G(t)$. Here, $B(t)$ is a $(k \times q)$ matrix that converges to a rank k matrix B and $G(t)$ is a $(q \times 1)$ mean-zero orthogonality condition used to estimate β. $G(t)$ is an average of a random variable $g(t)$ that has expectation 0; the average is taken over 1 to R (fixed scheme), $t-R+1$ to t (rolling scheme), or 1 to t (recursive scheme). When maximum likelihood is used, $q = k$, $B(t)$ is the Hessian evaluated on the line between $\hat{\beta}(t)$ and β, and $G(t)$ is the score. When instrumental variables (more generally, generalized method of moments) is used, $q \ge k$ ($q > k$ in overidentified models) and $B(t)$ is a matrix selecting which combination of orthogonality conditions to set to 0. When the models are nonnested, as is usually the case in practice, $B(t)$ and its large-sample counterpart B are block diagonal, and $G(t)$ simply stacks the orthogonality conditions from the two models. (The results here allow the two models to share some orthogonality conditions, as will also often be the case in practice.)

Write the forecast error (which may be multi- rather than one-step-ahead) as $e_{it} = e_{it}(\beta_i)$, with sample counterpart $\hat{e}_{it}$. Define

$$
\begin{aligned}
&\underset{(1\times 1)}{S} \equiv \sum_{j=-\infty}^{\infty} E(d_t - Ed_t)(d_{t-j} - Ed_t),\\
&\underset{(q\times q)}{S_{gg}} \equiv \sum_{j=-\infty}^{\infty} Eg_t g_{t-j}',\\
&\underset{(1\times q)}{S_{dg}} \equiv \sum_{j=-\infty}^{\infty} E(d_t - Ed_t)g_{t-j}', \quad \underset{(k\times k)}{V_\beta} \equiv BS_{gg}B',\\
&\underset{(1\times 1)}{d_t} \equiv d_t(\beta) \equiv [e_{1t}(\beta_1)]^2 - e_{1t}(\beta_1)e_{2t}(\beta_2),\\
&\underset{(1\times k)}{D} \equiv E[\partial d_t/\partial\beta]',\\
&\underset{(1\times 1)}{d} \equiv n^{-1}\Sigma\hat{d}_t = n^{-1}\Sigma(\hat{e}_{1t}^2 - \hat{e}_{1t}\hat{e}_{2t}).
\end{aligned}
$$

V_β is the asymptotic variance–covariance matrix of the estimator of β. Then

$$
\begin{aligned}
\sqrt{n}(\bar{d} - Ed_t) &\sim_A \mathrm{N}(0, \Omega),\\
\Omega &\equiv S + \lambda_{dh}(DBS_{dg}' + S_{dg}B'D')\\
&\quad + \lambda_{hh}DV_\beta D'. \qquad \text{(A.2)}
\end{aligned}
$$

The obvious sample analogues may be used in estimation of the quantities in (A.2). Autocovariances of $\hat{d}_t$ and $\hat{g}_t \equiv g_t(\hat{\beta}_t)$ may be used to estimate S, S_{gg}, and S_{dg}, using nonparametric kernel estimators such as those of Andrews (1991) and Newey and West (1994) if desired; standard estimators of regression variance–covariance matrices may be used to estimate V_β; a sample average of $\partial\hat{d}_t/\partial\beta$ may be used to estimate D.

Under the null of encompassing ($Ed_t = 0$), simplifications usually obtain. Let me illustrate first with linear models $y_t = X_{1t}'g_1(\beta_1) + e_{1t}$ and $y_t = X_{2t}'g_2(\beta_2) + e_{2t}$. [To clarify the notation: If (say) Model (1) is an autoregression of order $k_1 - 1$, then X_{1t} includes a constant and $y_{t-h}, y_{t-h-1}, \ldots, y_{t-h-k1+2}$; β_1 includes the autoregressive parameters; $X_{1t}'g_1(\beta_1)$ is the usual h-step-ahead autoregressive forecast.] A sufficient condition for $Ed_t = 0$ is

$E(e_{1t}|X_{1t}, X_{2t}, y_{t-h}, X_{1t-1}, X_{2t-1}, y_{t-h-1}, \dots) = 0$. [If Model (1) is an autoregression, then of course it is redundant to include lagged y_t's along with lagged X_{1t}'s in the conditional expectation.] In the more general, possibly nonlinear case, sufficient conditions for $Ed_t = 0$ are that e_{1t} has mean 0 conditional on (1) y_{t-j} for $j \geq h$, (2) current and lagged values of predictions from each of the models, and (3) current and lagged values of the derivatives of e_{1t} and e_{2t} with respect to β_1 and β_2.

Under these conditions, the autocorrelations of d_t will be 0 after the hth, implying $S = \sum_{j=-h+1}^{h-1} Ed_t d_{t-j}$. Similarly, D will have the form $[-Ee_{2t}(\partial e_{1t}/\partial\beta_1), 0_{k2\times 1}]$. This means that $DV_\beta D'$ will reduce to $[Ee_{2t}(\partial e_{1t}/\partial\beta_1)]'V_{\beta 1}[Ee_{2t}(\partial e_{1t}/\partial\beta_1]$, where $V_{\beta 1}$ is the $k_1 \times k_1$ variance–covariance matrix of β_1, and there will be no need to explicitly calculate the cross-covariance between the estimators of β_1 and β_2. The middle term in (A.2), $DBS'_{dg} + S_{dg}B'D'$, simplifies analogously.

If neither model encompasses the other, more complex calculations are required. See West (in press).

[Received July 1999. Revised April 2000.]

REFERENCES

Andrews, D. W. K. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 1465–1471.

Chong, Y. Y., and Hendry, D. F. (1986), "Econometric Evaluation of Linear Macro-Economic Models," *Review of Economic Studies*, 53, 671–690.

Cooper, R. L. (1972), "The Predictive Performance of Quarterly Econometric Models of the United States," in *Econometric Models of Cyclical Behavior, Volume II*, ed. B. Hickman, New York: Columbia University Press, pp. 813–925.

Diebold, F. X., and Mariano, R. S. (1995), "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–263.

Engle, R. F., Hong, C. and Kane, A. (1990), "Valuation of Variance Forecasts With Simulated Options Markets," Working Paper 3350, National Bureau of Economic Research, Cambridge, MA.

Ericcson, N. R., and Marquez, J. (1993), "Encompassing the Forecasts of U.S. Trade Balance Models," *Review of Economics and Statistics*, 75, 19–31.

Harvey, D. I., Leybourne, S. J., and Newbold, P. (1998), "Tests for Forecast Encompassing," *Journal of Business & Economic Statistics*, 16, 254–259.

Newey, W. K., and West, K. D. (1994), "Automatic Lag Selection in Covariance Matrix Estimation," *Review of Economic Studies*, 61, 631–654.

West, K. D. (1996), "Asymptotic Inference About Predictive Ability," *Econometrica*, 64, 1067–1084.

——— (in press), "Encompassing Tests When No Model Is Encompassing," *Journal of Econometrics*, 97.

West, K. D., and McCracken, M. W. (1998), "Regression Based Tests of Predictive Ability," *International Economic Review*, 39, 817–840.