ASYMPTOTIC INference ABOUT PREDICTIVE ABILITY

BY KENNETH D. WEST

This paper develops procedures for inference about the moments of smooth functions of out-of-sample predictions and prediction errors, when there is a long time series of predictions and realizations. The aim is to provide tools for analysis of predictive accuracy and efficiency, and, more generally, of predictive ability. The paper allows for nonnested and nonlinear models, as well as for possible dependence of predictions and prediction errors on estimated regression parameters. Simulations indicate that the procedures can work well in samples of size typically available.

KEYWORDS: Forecasting, forecast evaluation, testing, hypothesis test, model comparison.

1. INTRODUCTION

This paper develops procedures for asymptotic inference about the moments of smooth functions of predictions and out-of-sample prediction errors. The relevant environments are ones in which a long time series of predictions has been made from a sequence of base periods, and, if predictions are based on regression estimates, each of the sequence of regression estimates underlying the predictions has also been obtained from a long time series.

The aim is to provide tools that will be useful in two broad classes of applications. The first evaluates a model or a set of models on the basis of predictive accuracy or efficiency. Common measures of accuracy or efficiency include mean or mean squared prediction error, correlation between one model’s prediction and another model’s prediction error, correlation between prediction and realization, and serial correlation of prediction errors (e.g., Nelson (1972), Meese and Rogoff (1983), Fair and Shiller (1990), Stock and Watson (1993), West and Cho (1995)). The second, closely related, class of applications uses a series of predictions to make hypothetical decisions about, say, asset allocation, and then measures the quality of a model by the mean profit or utility yielded by use of its predictions (e.g., McCulloch and Rossi (1990)).

Diebold and Mariano (1994) provide a review and extension of procedures to perform inference about predictions when predictions do not rely on regression estimates. Typically, however, economic predictions rely on such estimates.

1I thank two anonymous referees, the Co-Editor, Patrick DeFontnouvelle, Steven Durlauf, Whitney Newey, Mark Watson, Ka-Fu Wong, Jeff Wooldridge, and seminar participants at the Federal Reserve Board of Governors, NBER, Northwestern University, and the Universities of Chicago, Iowa, Rochester, and Wisconsin for helpful comments and discussions. I also thank Chia-Yang Hueng, John Jones, Michael McCracken, and Ka-Fu Wong for excellent research assistance, and the NSF, the Sloan Foundation, and the University of Wisconsin Graduate School and College of Letters and Science for financial support.
Asymptotic tests that examine the correlation between realizations or prediction errors on the one hand and regression-based predictions on the other have been applied by a number of authors, one of the earliest being Nelson (1972). For asymptotic inference about mean squared prediction errors, Meese and Rogoff (1988), Christiano (1989), and Diebold and Rudebusch (1991) suggest applying standard asymptotic results (e.g., Hansen (1982)) in a fashion that, in the end, entails ignoring possible effects of error in estimation of parameters. These papers do not, however, justify their procedures with formal theory or with simulation evidence.

Simulation methods to compare the predictive accuracy of competing models have been proposed by Fair (1980). Such methods tend, however, to be computationally intensive, and may require specification of a null model to generate data, a specification that sometimes is not easy to come by when one is comparing competing, non-nested models. Nevertheless, simulation methods may be applied quite generally, and the asymptotic procedures proposed here should be viewed as complementary rather than competing.

These procedures allow for: a variety of linear and nonlinear techniques to estimate the models used to make the predictions, including maximum likelihood and generalized method of moments; serial correlation and conditional heteroskedasticity in regression disturbances and prediction errors; comparison of non-nested models; inference about moments of general, nonlinear functions of single or multi-period predictions and prediction errors. Using standard regularity conditions, I establish consistency and asymptotic normality of the estimators of the moments, and show that the asymptotic variance-covariance matrix may be estimated by familiar methods, including kernel techniques that allow for unknown forms of serial correlation and heteroskedasticity in prediction errors. A small set of simulations suggests that the procedure can yield accurately sized tests in samples of size typically available.

Whether and how one adjusts for error in estimation of regression parameters used to form predictions depends on a number of factors: the moment being examined, the regression technique, the fraction of the total sample used for out-of-sample estimation of the moments, and, of course, the probabilistic environment. Simulation evidence indicates that substantial size distortions may result if one ignores uncertainty about the regression vector that, according to this paper's asymptotic approximation, is asymptotically relevant for inference.2

Section 2 describes the basic environment, Section 3 presents technical assumptions, Section 4 derives asymptotic results, Section 5 presents Monte

---

2It should be acknowledged at the outset that the paper does not provide a formal statistical justification for the use of out-of-sample rather than in-sample analysis: the conventional statistical theory used here indicates that a model that does not fit well in-sample will not fit well out-of-sample. But out-of-sample comparisons sometimes bring surprising and important insights (e.g., Nelson (1972), Meese and Rogoff (1983)), perhaps because of inadvertent over-fitting that results from repeated profession-wide use of a limited body of data. This suggests that tests of predictive ability have good power, and, in any case, the tools developed here are relevant for work aimed at developing good methods to predict out-of-sample (e.g., Stock and Watson (1993)).
Carlo evidence, and Section 6 concludes. All proofs are in an Appendix. Additional proofs, simulation results, discussion, and interpretation have been omitted from the paper to save space, but are available in an additional Appendix and in two earlier versions of this paper that are collectively referenced as West (1994).

2. DESCRIPTION OF ENVIRONMENT

One is interested in an \((l \times 1)\) vector of moments \(E f_t\), where \(f_t\) depends as described below on the \((k \times 1)\) unknown parameter vector \(\beta^*\), and the \(t\) subscript indicates that \(f(\cdot)\) depends not only on \(\beta^*\) but also on observable data. If estimated serial correlation coefficients are used in prediction, these will be included along with regression parameters in \(\beta^*\). If moments of predictions or prediction errors of competing sets of regressions are to be compared, the parameter vectors from the various regressions are stacked to form \(\beta^*\). A special case of the analysis below is when predictions do not depend on estimated parameters, so that \(k = 0\); to avoid repeated qualifications to statements concerning ranks of certain matrices, I assume \(k > 0\) and note in Section 4 how the results specialize when \(k = 0\).

Let \(\tau \geq 1\) be the longest prediction horizon of interest. There are \(P\) predictions or vectors of predictions in all. The first prediction is based on a parameter vector estimated using data from 1 to \(R\) (\(R\) as in Regress), the second on a parameter vector estimated using data from 1 to \(R + 1\), \ldots, the last on a parameter vector estimated using data from 1 to \(R + P - 1\) \(= T\). \(^3\) Thus, \(R + P - 1 + \tau = T + \tau\) is the size of available sample. Division of the available data into \(R\), \(P\), and \(\tau\) is taken as given; when a model of interest was first estimated on data from \(t = 1, \ldots, R\), there was a natural choice of \(R\).

Let \(\hat{\beta}_t\) denote an estimate of \(\beta^*\) that relies on data from period \(t\) and earlier. The estimate of \(E f_t\) is constructed as

\[
\hat{f} = P^{-1} \sum_{t=R}^{T} f_{t+\tau}(\hat{\beta}_t).
\]

To illustrate \(f_t\), consider the scalar linear model \(y_t = X'_t \beta^* + u'_t\), and assume that the population one-step-ahead prediction of \(y_{t+1}\) is computed as \(X'_{t+1} \beta^*\). The corresponding sample prediction would be \(X'_{t+1} \hat{\beta}_t\), where, for example, \(\hat{\beta}_t = (\sum_{s=1}^{t} X_s X'_s)^{-1} \sum_{s=1}^{t} X_s y_s\) if ordinary least squares is the estimator.

If the object of interest is the one-step-ahead mean squared prediction error (MSPE), then \(E f_t = E(y_t - X'_t \beta^*)^2\) and \(\hat{f} = P^{-1} \sum_{t=R}^{T} (y_{t+1} - X'_{t+1} \hat{\beta}_t)^2\). If the object of interest is one-step-ahead mean prediction error (MPE), then \(E f_t = E(y_t - X'_t \beta^*)\), and here, and in all subsequent examples but the last, \(\hat{f}\) is constructed in the obvious way. If, further, \(X_{t+1} = y_t\), so that \(k = 1\) and the

\(^3\)West (1994) considers how the results vary if instead: (a) a single estimate of \(\beta^*\) is used to form all \(P\) out of sample predictions, or (b) the sequence of \(P\) estimates of \(\beta^*\) is obtained from a series of rolling samples, each of size \(R\).
model is a zero mean univariate autoregression, the $\tau$-step-ahead MSPE is $E[y_{t+\tau+1} - X_{t+\tau+1}'(\beta^*)]$. As indicated in this example, the dating of $X_{t+1}$ is arbitrary, and $X_{t+1}$ may be realized in period $t$ or earlier. Finally, for the first autocovariance of $u_t$, $E(y_{t+1} - X_{t+1}'\beta^*)(y_t - X_t'\beta)$, we have $\hat{f} = P^{-1} \sum_{t=R} y_{t+1} - X_{t+1}'\hat{\beta}(y_t - X_t'\hat{\beta})$. (The natural, and conventional, way to compute the corresponding sample moment is $P^{-1} \sum_{t=R} y_{t+1} - X_{t+1}'\hat{\beta}(y_t - X_t'\hat{\beta})$, which is asymptotically equivalent (West (1994)). To keep the notation relatively simple, in the theory I assume $\hat{f}$ depends only on $\hat{\beta}_t$, although I use the natural estimator in the simulations.)

It will rarely be the case in practice that $l = 1$, as in these examples. But more realistic examples may be built from such simple ones by stacking the objects of interest into a vector $f_l$ with $l > 1$. See, for example, the experiment described in Section 5.2 below.

3. ASSUMPTIONS

The first assumption is that $f_t$ is well approximated by a smooth quadratic in the neighborhood of the parameter vector. I use the following notation: for any differentiable function $g_t: R^m \to R^s$ and for $x$ in the domain of $g_t$, let $\partial g_t / \partial x$ denote the $(s \times m)$ matrix of partial derivatives of $g_t$; for any matrix $A = [a_{ij}]$, let $|A| = \max_{i,j}|a_{ij}|$; summations of variables indexed by $t$ or $t + \tau$ run from $t = R$ to $t = T = R + \tau - 1$: for any variable $x$, $\sum x(t) = \sum_{t=R}^T x(t)$, $\sum x_{t+\tau} = \sum_{t=R}^T x_{t+\tau}$; summations of variables indexed by $s$ run from $1$ to $t$; for any variable $x$, $\sum x_s = \sum_{s=1}^t x_s$.

ASSUMPTION 1: In some open neighborhood $N$ around $\beta^*$, and with probability one:

(a) $f_t(\beta)$ is measurable and twice continuously differentiable with respect to $\beta$.

(b) Let $f_{it}$ be the $i$th element of $f_t$. For $i = 1, \ldots, l$ there is a constant $D < \infty$ such that for all $t$, $\sup_{\beta \in N} |\partial^2 f_{it}(\beta) / \partial \beta \partial \beta'| < m_t$ for a measurable $m_t$ for which $Em_t < D$.

Provided appropriate second moments exist, Assumption 1 holds trivially when the function is MPE, MSPE, or an autocovariance, the horizon is one period, and the model is linear. It also holds when these functions are based on: (a) multiperiod ARIMA forecasts and single or multiperiod reduced form forecasts for linear simultaneous equations models, all of which are nonlinear in the parameters; (b) predictions based on smooth transformations of forecasts constructed from regression estimates (e.g., Fair and Shiller (1990), West et al. (1993)).

ASSUMPTION 2: The estimate $\hat{\beta}_t$ satisfies $\hat{\beta}_t - \beta^* = B(t)H(t)$, where $B(t)$ is $(k \times q)$ and $H(t)$ is $(q \times 1)$, with (a) $B(t)^\dagger B$, $B$ a matrix of rank $k$; (b) $H(t) = t^{-1} \sum_{s=1}^t h_s(\beta^*)$ for a $(q \times 1)$ orthogonality condition $h_s(\beta^*)$; (c) $Eh_s(\beta^*) = 0$. 
\( H(t) \) is a sample average of a \((q \times 1)\) orthogonality condition used to identify \( \beta^* \), \( B(t) \) a \((k \times q)\) matrix that selects a linear combination of these sample averages.

To illustrate, let \( h_t = h_t(\beta^*) \), and assume stationarity. If, say, the only regression equation is the scalar model \( y_t = X_t'\beta^* + u_t \), and the regression technique is OLS, \( B = (EX_tX'_t)^{-1} \) and \( h_t = X_tu_t \). Note that there is no presumption that \( u_t \) or \( X_tu_t \) is serially uncorrelated (although it is of course required that \( EX_tu_t = 0 \)). Suppose more generally that for \( y_t \) \((n \times 1)\) and a smooth function \( g : R^k \to R^n \), the set of regression equations is \( y_t = g(X_t, \beta^*) + u_t \). Let \( \beta^* \) be estimated by GMM: \( \beta_t \) is chosen to minimize \( \sum h_t W(t) \sum h_t \), \( h_t = [y_t - g(X_t, \beta_t)] \otimes Z_t, Z_t \) a vector of instruments, \( Eh_t = Eu_t \otimes Z_t = 0 \), and \( W(t) \geq W \) p.d. and symmetric. Then

\[
B(t) = - \left[ A(t)^{-1} \sum \frac{\partial h_t}{\partial \beta} \right]^{-1} A(t), \quad \text{where}
\]

\[
A(t) = \left[ -t^{-1} \sum \frac{\partial h_t}{\partial \beta} (\hat{\beta}_t) \right]' W(t),
\]

\[- \frac{\partial h_t}{\partial \beta} = \frac{\partial g_t}{\partial \beta} \otimes Z_t, \]

and \( \partial h_t / \partial \beta \) is \( \partial h_t / \partial \beta \) evaluated at points between \( \hat{\beta}_t \) and \( \beta^* \); optimal GMM estimation (again, allowed but not required) entails \( W = \sum \rightarrow \infty Eh_t h'_t \). The analogous definition for maximum likelihood is that \( B(t) \) is the inverse of the Hessian evaluated at a vector between the population parameter vector and its estimate, and \( H(t) \) is the score. See Hansen (1982) or Andrews (1987) for primitive conditions that imply Assumption 2.

Let

\[ f_t = f_t(\beta^*), \quad f_{1\beta} = \frac{\partial f_t}{\partial \beta}(\beta^*), \quad F = Ef_{1\beta}. \]

**Assumption 3:** (a) For some \( d > 1 \), sup \( E[||\text{vec}(f_{1\beta})|^4, (f_t' - F)^4, \left| f_t' - F \right|^4 d \rightarrow \infty \), where \( || \cdot || \) denote Euclidean norm. (b) \( [\text{vec}(f_{1\beta} - F), (f_t - F)^4, h_t, h_t'] \) is strong mixing, with mixing coefficients of size \( -3d/(d - 1) \). (c) \( [\text{vec}(f_{1\beta} Y, f_t', h_t')] \) is covariance stationary. (d) Let \( \Gamma_{ff}(j) = Ef_t (f_t' - Ef_t)(f_t - Ef_t)' \), \( S_{ff} = \sum \rightarrow \infty \Gamma_{ff}(j) \). Then \( S_{ff} \) is p.d.

Assumption 3 allows serial correlation and conditional heteroskedasticity in both \( f_t - Ef_t \) and \( h_t \). The assumption is rather stronger than is necessary, but is convenient.

**Assumption 4:** \( R, P \rightarrow \infty \) as \( T \rightarrow \infty \), and \( \lim_{T \rightarrow \infty} (P/R) = \pi \), \( 0 \leq \pi \leq \infty \); \( \pi = \infty \Leftrightarrow \lim_{T \rightarrow \infty} (R/P) = 0 \).
\( R \to \infty \) allows use of asymptotic theory to formalize how uncertainty about \( \beta^* \) affects the estimated moment; even in the absence of such uncertainty, \( P \to \infty \) is obviously necessary to allow asymptotic analysis. The parameter \( \pi \) is taken as given. Given that the forecast horizon \( \tau \) is fixed, the decision to design a symbol to the limit of \( P/R \) rather than of, say, \( P/(T + \tau) \) is arbitrary. That \( \tau \) is fixed also suggests that the relevant applications are those in which \( \tau \ll R, P \).

Throughout, I maintain Assumptions 1–4.

4. ASYMPTOTIC DISTRIBUTION OF THE OUT-OF-SAMPLE MEAN OF \( f_t \)

The Appendix applies a mean value expansion of \( f_t + \tau(\hat{\beta}_t) \) around \( \beta^* \) to show

\[
P^{1/2}(\hat{f} - Ef_t) = P^{-1/2} \sum (f_t + \tau - Ef_t) + FB[P^{-1/2} \sum H(t)] + o_p(1).
\]

The first term on the right-hand side represents uncertainty about \( \hat{f} \) that is present even when \( \beta^* \) is known. The second term represents uncertainty about \( \beta^* \). To state precisely how the asymptotic variance of \( P^{1/2}(\hat{f} - Ef_t) \) reflects both forms of uncertainty requires some additional notation. Let

\[
\begin{align*}
\Gamma_{fh}(j) &= E(f_t - Ef_t)k_{t-j}, \quad S_{fh} = \sum_{j=-\infty}^{\infty} \Gamma_{fh}(j), \quad \Gamma_{hh}(j) = Eh_t k_{t-j}, \\
S_{hh} &= \sum_{j=-\infty}^{\infty} \Gamma_{hh}(j), \quad S = \begin{pmatrix} S_{ff} & S_{fh}B' \\ BS_{fh}' & BS_{hh}B' \end{pmatrix} = \begin{pmatrix} S_{ff} & S_{fh}B' \\ BS_{fh}' & V_{\beta} \end{pmatrix}, \\
\Pi &= \Pi(\pi) = 1 - \pi^{-1}\ln(1 + \pi) \quad \text{for} \quad 0 < \pi < \infty, \\
\Pi &= 0 \quad \text{for} \quad \pi = 0, \quad \Pi = 1 \quad \text{for} \quad \pi = \infty.
\end{align*}
\]

Apart from a scale factor, \( S \) is the \((l+k) \times (l+k)\) spectral density of \([f_t - Ef_t, h_t B']^\prime\) at frequency zero. The \((k \times k)\) matrix \( V_{\beta} = BS_{hh}B' \) is the asymptotic variance-covariance matrix of \( T^{1/2}(\hat{\beta}_T - \beta^*) \). The scalar function \( \Pi \) is discussed in comment 4 below.

**Lemma 4.1:**
(a) \( P^{-1/2} \sum (f_t + \tau - Ef_t) \xrightarrow{d} N(0, S_{ff}), \quad E[P^{-1} \sum H(t) \sum H(t)'] \to 2 \Pi S_{hh}, \quad E[P^{-1} \sum (f_t + \tau - Ef_t) \sum H(t)'] \to \Pi S_{fh}. \)
(b) If \( S \) is p.d., \( P^{1/2} \sum (f_t + \tau - Ef_t, \sum H(t) B' \sum H(t) \sum H(t)') \xrightarrow{d} N(0, V) \), where

\[
V = \begin{pmatrix} S_{ff} & \Pi S_{fh} B' \\
\Pi BS_{fh}' & 2 \Pi V_{\beta} \end{pmatrix};
\]

\( V \) has full rank if \( \pi \neq 0 \).

**Theorem 4.1:**
(a) If \( \pi = 0 \) or \( F = 0 \), \( P^{1/2}(\hat{f} - Ef_t) \xrightarrow{d} N(0, \Omega), \quad \Omega = S_{ff}. \)
(b) If \( S \) is p.d., \( P^{1/2}(\hat{f} - Ef_t) \xrightarrow{d} N(0, \Omega), \quad \Omega = S_{ff} + \Pi(FBS_{fh} + S_{fh} B' F') + 2 \Pi FV_{\beta} F'. \)
The two parts of Theorem 4.1 of course are not mutually exclusive; the formula for the rank l matrix $\Omega$ in Theorem 4.1(b) reduces to $S_{ff}$ under the conditions of Theorem 4.1(a). Primitive conditions that ensure that $\hat{S}$ is p.d. may be found in, for example, Durbin (1970) when $E_{f}$ equals the first autocovariance of a disturbance, and Hamilton (1994, p. 301) when $E_{f} = MSPE$ in a vector autoregression. Consistency ($\tilde{\beta} \overset{(d)}{\to} E_{f}$) obviously follows from Theorem 4.1. If $\beta^*$ is known (no regression estimation is required for prediction), $B = 0$ and, as noted by Diebold and Mariano (1994), $P^{1/2}(\hat{\beta} - E_{f}) \overset{(d)}{=} N(0, S_{ff})$.

The following six comments on Theorem 4.1 are relevant. The first three and the fifth discuss cases in which $P^{1/2}(\hat{\beta} - E_{f}) \overset{(d)}{=} N(0, S_{ff})$ and uncertainty about $\beta^*$ is asymptotically irrelevant for inference about $E_{f}$; I make a point of noting these because in such cases inference is greatly simplified.

1. From part (a) of Theorem 4.1, one condition for asymptotic irrelevance is when $0 = \pi = \lim_{T \to \infty}(P/R)$. The intuition is that one may treat the estimate of $\beta^*$ as known, if the asymptotic distribution of $\tilde{\beta}$ is derived under the assumption that for arbitrarily large $T$ there will be an arbitrarily large number of observations used for estimation of $\beta^*$ relative to the number used for estimation of $E_{f}$. This point was argued informally by Chong and Hendry (1986) in the context of tests of forecast encompassing.

2. Theorem 4.1(a) also indicates that asymptotic irrelevance holds when $0 = F = E[(\partial f_i/\partial \beta)(\beta^*)]$. The leading case here is MSPE when the predictors are uncorrelated with the prediction error. In the scalar linear model $y_i = X_i^t \beta^* + u_i$ with one step ahead population prediction $X_{i+1}^t \beta^*$; for example, one has $f_i(\beta^*) = (y_i - X_i^t \beta^*)^2$.

$$\frac{\partial f_i}{\partial \beta}(\beta^*) = -2u_iX_i^t \Rightarrow F = -2Eu_iX_i^t \quad \text{and} \quad F = 0 \quad \text{if} \quad Eu_iX_i^t = 0.$$

More generally, asymptotic irrelevance applies for the MSPE of single or multiperiod predictions from, for example, the reduced form of simultaneous equations models. When one is comparing non-nested models, the appropriate asymptotic variance is still $S_{ff}$ as long as each prediction error is uncorrelated with the corresponding predictors, even if the competing models use different information sets so that one model’s prediction error is correlated with another model’s prediction.

3. On the other hand, asymptotic irrelevance sometimes applies to an out-of-sample test when it does not apply to the corresponding in-sample test. The technical condition here is that $\Pi(FB_{fh}^t + S_{fh}B^tF^t) + 2\Pi FV_{\beta}F^t = 0$, so that the variance induced by error in estimation of $\beta^*$ is exactly offset by the covariance between such terms and terms that would be present even if $\beta^*$ were known.

One such case is in testing for first order serial correlation in one step ahead prediction errors in certain models. To illustrate, consider once again the scalar linear model, now specialized for simplicity so that $X_i$ is a scalar and $X_i = y_{i-1} \Rightarrow y_i = y_{i-1} \beta^* + u_i$. Assume $E(u_i | y_1, y_{-2}, \ldots) = 0$ and let the object of inter-
est be $\rho = (Eu_i^2)^{-1}Eu_iu_{i-1} = \gamma_0^{-1}\gamma_1 = \gamma_0^{-1}Ef_i (= 0)$. Let the estimator be OLS, let $\hat{u}_i$ be the one-step-ahead prediction error, $\hat{\gamma}_1 = P^{-1} \Sigma \hat{u}_i \hat{u}_{i-1}^{-1} \hat{f} = \hat{\gamma}_0$, $\hat{\rho} = \hat{\gamma}_0^{-1} \hat{f}$. Theorem 4.1 indicates that the asymptotic variance of $P_1^{1/2} \hat{\rho}$ is

$$
(4.3) \quad \gamma_0^{-2} \Omega = \gamma_0^{-2} \left( Eu_i^2 u_{i-1}^2 - 2 \Pi \left[t(Eu_i, y_i)(Ey_i^2)^{-1} Eu_i^2 u_{i-1} y_{i-1} \right] \right) + 2 \Pi \left[t(Eu_i, y_i)(Ey_i^2)^{-2} (Ey_i^2 u_i^2) \right].
$$

It may be seen that if $E(u_i^2|y_{i-1}, y_{i-2}, \ldots) = Eu_i^2 = \gamma_0$, the right-hand side of (4.3) reduces to $\gamma_0^{-2} = 1$: the standard error to be used in hypothesis testing is the familiar $1/P^{1/2}$ that applies as well when $\beta^*$ is known. In contrast, it is well known (e.g., Durbin (1970)) that if in-sample residuals are used, the relevant asymptotic variance includes a term reflecting uncertainty about $\beta^*$. The result illustrated in (4.3) holds generally when predicting from the reduced form of a linear simultaneous equations model estimated by 3SLS (a VAR estimated by OLS is a special case), as long as the disturbance vector is a conditionally homoskedastic martingale difference.

4. Such asymptotic irrelevance of course does not always apply. One example is MSPE when the predictors are correlated with the prediction error (see Section 5 below). Also, in the tests for serial correlation and for zero covariance just described in comment 3, asymptotic irrelevance typically fails if the disturbance is conditionally heteroskedastic and there are lagged dependent variables, even if the disturbance is a martingale difference. This may be verified by working through (4.3) with a specific parametric model for the conditional heteroskedasticity.

To interpret Theorem 4.1(b) in these and other cases, note first that $0 \leq \Pi \leq 1$ and $\Pi'(\pi) > 0$. Uncertainty about $\beta^*$ adds $\Pi\{FBS_{fh} + S_{fh}B'F'\} + 2\Pi F\beta F'$ to the asymptotic variance of $P^{1/2}(\hat{f} - Ef_i)$. In comparison to the hypothetical case when $\beta^*$ is known, uncertainty about $\beta^*$ may cause an asymptotic standard error to be larger, smaller, or, as noted in comment 3, no different. In those cases when uncertainty about $\beta^*$ causes the standard error of interest to be larger (smaller), the standard error increases (decreases) with the fraction of the sample set aside for prediction, that is, with $\pi$.

5. This comment concerns MPE. Theorem 4.1 may not be applicable because $S$ may not be p.d. despite positive definiteness of $S_{ff}$ and of $V_{\beta}$. West (1994) shows, however, it is generally the case that a vector of mean prediction errors is asymptotically normal with variance-covariance matrix $S_{ff} = \sum_{i=-\infty}^{\infty} Eu_i u_{i-1}$ in the model $y_i = X_i' \beta^* + u_i$. In contrast to the serial correlation example (comment 3), such asymptotic irrelevance holds even if the disturbance vector is conditionally heteroskedastic and serially correlated.

6. The final comment concerns consistent estimation of $\Omega$. Inspection of Theorem 4.1 indicates that one must estimate $S_{ff}$ and, when uncertainty about $\beta^*$ matters, one must also supply a value for $\Pi$ and estimate $B$, $F$, $S_{fh}$, and $S_{sh}$. Sample analogues for the population quantities may be used (West (1994)): One
can set

\[ \Pi = 1 - \left( \frac{P}{R} \right)^{-1} \ln \left( 1 + \frac{P}{R} \right). \]

By assumption 2, \( \hat{B} \equiv B(T) \xrightarrow{p} B \). Next,

\[ \hat{F} = P^{-1} \sum_{t=R}^{T} \frac{\partial f_{t+\tau}}{\partial \beta} (\hat{\beta}_t) \xrightarrow{p} F \]

and sample autocovariances of \( f_t \) and \( h_t \) are consistent (e.g., letting \( \hat{f}_{t+\tau} = f_{t+\tau}(\hat{\beta}_t) \) and \( \hat{h}_t = h_t(\hat{\beta}_t) \), \( P^{-1} \sum_{t=R}^{T} (\hat{f}_{t+\tau} - \hat{F})(f_{t+\tau} - \hat{F})' \xrightarrow{p} \Gamma_{ff}(0) \), \( P^{-1} \sum_{t=R}^{T} (\hat{f}_{t+\tau} - \hat{F}) \hat{h}_t \xrightarrow{p} \Gamma_{fh}(0) \), \( P^{-1} \sum_{t=R}^{T} \hat{h}_t (\hat{f}_{t+\tau} - \hat{F})' \xrightarrow{p} \Gamma_{hh}(0) \), with analogous results for auto- and cross-covariances at nonzero lags). Often, this suffices for inference, as in the simulations below. But if not, upon mild strengthening of the assumptions about \( h_t \) along the lines of Andrews (1991), conventional nonparametric kernel estimators, such as the Bartlett or quadratic spectral, may be used. When applied to the sample auto- and cross-covariances of \( f_t \) and \( h_t \), these yield estimates \( \hat{S}_{ff} \), \( \hat{S}_{fh} \), and \( \hat{S}_{hh} \) that are consistent for \( S_{ff} \), \( S_{fh} \), and \( S_{hh} \), and, as well, a consistent p.s.d. estimate \( \hat{\Omega} \equiv \hat{S}_{ff} + \Pi(\hat{F}\hat{B}\hat{S}_{fh} + \hat{S}_{fh}\hat{B}'\hat{F}') + 2\Pi \hat{F}\hat{B}\hat{S}_{hh}\hat{B}'\hat{F}'. \)

5. MONTE CARLO EVIDENCE

Here I present two small Monte Carlo experiments, aimed at giving a feel for whether Theorem 4.1’s asymptotic approximation might yield well-sized test statistics. Both involve tests on various moments of one-step-ahead prediction errors in models estimated by instrumental variables. The first does not require accounting for uncertainty about the regression vector, the second does.

In the experiments, each of 5000 artificial samples of size 200 was split into 12 different regression (\( R \)) and prediction (\( P \)) samples: \( R = 25, P = 25, 50, 100, 150, 175; R = 50, P = 25, 50, 100, 150; R = 100, P = 25, 50, 100 \). This range for \( P/R \) (from 25 to 7), as well as the values of \( T = P + R - 1 \), seem broad enough to include most relevant empirical work. For each pair of \( R \) and \( P \), the first \( R + P \) observations of each sample of size 200 were used. So \( R = 50/P = 100 \) and \( R = 100/P = 50 \), for example, used the same 150 observations, but began the out-of-sample exercise at different points. I report tests of nominal size .05. Tests of nominal size .01 and .10 worked equally well (West (1994)).

5.1. Predictions Using Predetermined Variables in a Simultaneous Equations Model

Here I consider a bivariate model, borrowed from Schmidt (1977):

\begin{align*}
 y_{1t} &= \beta_{11} + \beta_{12} y_{2t} + \beta_{13} y_{1t-1} + \beta_{14} x_{1t} + u_{1t} = X'_{1t} \beta_1 + u_{1t}, \\
 y_{2t} &= \beta_{21} + \beta_{22} y_{1t} + \beta_{23} y_{2t-1} + \beta_{24} x_{2t} + u_{2t} = X'_{2t} \beta_2 + u_{2t}. 
\end{align*}
In (5.1), \((u_{1t}, u_{2t})' = u_t\) is an iid normal vector. The exogenous variables \(x_{1t}\) and \(x_{2t}\) were assumed to follow univariate AR(1)'s with common parameter \(\phi\) and iid standard normal disturbances that are independent of one another and of \(u_t\). The \(5 \times 1\) vector of instruments is the list of predetermined variables in the reduced form, \(Z_t \equiv (1, y_{1t-1}, y_{2t-1}, x_{1t}, x_{2t})'\).

Most of the parameters were borrowed from Schmidt (1977): \(\beta^* = (\beta_1', \beta_2')' = (0, .2, .5, 1.0, 1.5, 2)'\); \(\phi = 0.8\); \(Eu_{1t}^2 = Eu_{2t}^2 = 1\); \(Eu_{1t}u_{2t} = 0.5\). For 5000 samples, an initial value of the \((4 \times 1)\) vector \((y_{10}, y_{20}, x_{10}, x_{20})'\) was drawn from the unconditional normal distribution of \((y_{1t}, y_{2t}, x_{1t}, x_{2t})'\), and samples of size 200 were generated. For each value of \(R\) and \(P\), and for \(t = R\) to \(t = R + P - 1\), 3SLS was used to estimate the structural equations (5.1) using the \(10 \times 1\) \((q = 10)\) orthogonality condition \(h_t = u_t \otimes Z_t\). The implied reduced form coefficients were computed, \(y_{1t+1}\) was predicted from the reduced form, and a \((P \times 1)\) vector of one step ahead prediction errors \(\hat{\theta}_{1t+1}\) was constructed.

This vector was used to perform the following three hypothesis tests: (a) Zero MPE: \(\hat{\beta} = P^{-1} \sum \hat{\theta}_{1t+1}, \hat{\Omega} = P^{-1} \sum \hat{\theta}_{1t}^2\); from comments 5 and 6 in Section 4, \(\hat{\beta} = P^{-1} \sum \hat{\theta}_{1t+1}, \hat{\Omega} = P^{-1} \sum \hat{\theta}_{1t}^2 - (P^{-1} \sum \hat{\theta}_{1t+1})^2\); from comments 2 and 6 in Section 4, \(\hat{\beta} = P^{-1} \sum \hat{\theta}_{1t+1}, \hat{\Omega} = P^{-1} \sum \hat{\theta}_{1t}^2 - (P^{-1} \sum \hat{\theta}_{1t}^2)^2\) (c) Zero first order serial correlation coefficient: the object of interest is \([P^{-1} \sum (\hat{\theta}_{1t}\hat{\theta}_{1t+1})]/(P^{-1} \sum \hat{\theta}_{1t}^2)\), and, in accordance with comment 3 in Section 4, the standard error used in hypothesis testing was simply \(P^{-1/2}\).

Table 1 has the results for nominal .05 tests. There is a mild tendency to overreject; of the 36 entries in the Table, only 8 are less than .05. (The entries are not, however, independent. If a test happens to overreject for a given \(R\) and \(P\), it is likely to do so for the same \(R\) and slightly larger \(P\) as well.) But the figures, which range between .034 and .100, are also close to .05.

5.2. MSPE with Endogenous Right-Hand-Side Variables as Predictors

Here I consider the system

\[
\begin{align*}
(5.2a) \quad y_t = \beta_1 + \beta_{12} w_{1t} + u_{1t} &= X_{1t}' \beta_1 + u_{1t}, \\
(5.2b) \quad y_t = \beta_2 + \beta_{22} w_{2t} + u_{2t} &= X_{2t}' \beta_2 + u_{2t},
\end{align*}
\]

where the \(w_{it}'s\) are correlated with the \(u_{it}'s\) and \(Eu_{1t}^2 = Eu_{2t}^2\). The idea is that (5.2a) and (5.2b) are competing models for \(y_t\). Write (5.2) as

\[
\begin{pmatrix}
y_t \\
y_{1t} \\
y_{2t}
\end{pmatrix} = Y_t = X_t' \beta^* + u_t =
\begin{pmatrix}
1 & w_{1t} & 0 & 0 \\
0 & 0 & 1 & w_{2t}
\end{pmatrix}
\begin{pmatrix}
\beta_1 \\
\beta_2 \\
u_{1t} \\
u_{2t}
\end{pmatrix}.
\]

Let \((z_{1t}, z_{2t}, v_t)'\) be an iid normal random vector with an identity covariance matrix. Let \(w_{1t} = z_{1t} + u_{1t}, w_{2t} = z_{2t} + v_{1t}\). The data are generated as \(y_t = w_{1t} + w_{2t} + u_{1t}\). So in (5.2a), \(\beta_{11} = \beta_{21} = 0, \beta_{12} = \beta_{22} = 1\). Each equation was estimated by 2SLS, with a vector of instruments \((1, z_{1t}, z_{1t}, Z_{1t} \text{ (equation (5.2a)) or } (1, z_{2t}, z_{2t}, Z_{2t}) \text{ (equation (5.2b))})
\]
### TABLE I

**Size of Nominal .05 Tests, Experiment 1**

<table>
<thead>
<tr>
<th></th>
<th>( R )</th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>150</th>
<th>175</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A. Mean Prediction Error</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>.059</td>
<td>.064</td>
<td>.060</td>
<td>.057</td>
<td>.060</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>.059</td>
<td>.063</td>
<td>.057</td>
<td>.059</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>.055</td>
<td>.058</td>
<td>.058</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>B. Mean Squared Prediction Error</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>.076</td>
<td>.066</td>
<td>.061</td>
<td>.063</td>
<td>.066</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>.081</td>
<td>.060</td>
<td>.057</td>
<td>.056</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>.100</td>
<td>.075</td>
<td>.055</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>C. First Order Serial Correlation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>.037</td>
<td>.046</td>
<td>.050</td>
<td>.049</td>
<td>.050</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>.038</td>
<td>.045</td>
<td>.053</td>
<td>.052</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>.034</td>
<td>.040</td>
<td>.043</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:**
1. As detailed in Section 5.1, the following was done 5000 times: a. 200 observations on the variables in a bivariate simultaneous equations model were generated. b. The model was estimated 175 times by 3SLS, using observations 1–25, 1–26, ..., and 1–199. The implied reduced form coefficients were used to make one-period-ahead predictions of one of the two endogenous variables. c. For the values of \( R \) and \( p \) given above, a \((P \times 1)\) vector of prediction errors was formed using predictions dated \( R, R + 1, \ldots, R + P - 1 \). d. The vector of sample prediction errors was used to test: zero mean (panel A), variance equal to the underlying population value (panel B), zero serial correlation (panel C).
2. According to this paper's theory, each statistic is asymptotically \( \chi^2(1) \). Each panel reports the fraction of the 5000 statistics that were greater than 3.84, which is the .05 critical value for a \( \chi^2(1) \) random variable.

\[ Z_{2s} (5.2b) \]. Let \( \hat{\beta}_{it} = (t^{-1} \sum_{s=1}^{t} Z_{is} X'_{is})^{-1}(t^{-1} \sum_{s=1}^{t} Z_{is} y_s) \) be the 2SLS estimate of \( \beta_i \). For each value of \( R \) and \( p \), and for \( t = R \) to \( T \), one step ahead prediction errors were obtained as \( \hat{\epsilon}_{it+1} = y_{it+1} - X'_{it+1} \hat{\beta}_{it} \). The hypothesis tested is that \( Eu_{1t}^2 = Eu_{2t}^2 \Rightarrow \hat{f} = (P^{-1} \sum \hat{\epsilon}_{it+1}^2, P^{-1} \sum \hat{\epsilon}_{2t+1}^2)^T \).

Note that realized right-hand-side endogenous variables are used for "prediction." In the forecasting literature such "conditional" or "ex-post" forecasts are made when one is not interested in ex-ante prediction but evaluation of predictive ability of a model given a path for some unmodelled (but endogenous, in a larger system) set of variables. Examples include Meese and Rogoff (1983, 1988) and Oliner et al. (1993). In such cases, \( F \) is nonzero and sampling error in estimation of \( \beta^* \) is asymptotically relevant (except when \( \pi = 0 \)).

In particular, in the system under consideration here,

\[ F = -2E \begin{pmatrix} u_{1t} & 0 \\ 0 & u_{2t} \end{pmatrix} X'_{it} = -2EU_{1t}X_{1t} \Rightarrow F(1,2) = -2Eu_{1t}w_{1t} \neq 0, \]

\[ F(2,4) = -2Eu_{2t}w_{2t} \neq 0. \]
Since \((u_{1i}, z_{1i}, u_{2i}, z_{2i})\) is normally distributed, \(S_{fh} = 0 \Rightarrow \Omega = S_{ff} + 8\Pi(EU_i'X_i')V_i(EX_i'U'_i)\). Thus use of \(S_{ff}\), which abstracts from sampling error in estimation of \(\beta^*\), will yield standard errors and confidence intervals that are too small, and will lead to too many rejections at any specified significance level. The lower the asymptotic ratio of observations used for prediction to those used for regression (the lower is \(\pi\)), the less will sampling error in estimation of \(\beta^*\) affect inference, and the closer will \(S_{ff}\) be to \(\Omega\).

\(\hat{\Omega}\) was constructed as follows. (i) In a computationally convenient variant of the estimator suggested in comment 6 in Section 4, \(V_{\beta} = B'S_{hh}B\) was estimated in conventional in-sample fashion, using the largest regression sample: For \(s = 1, \ldots, T\), let \(\tilde{u}_s = Y_s - X_s'\hat{\beta}_T\), \(\tilde{h}_s = (Z_{1s}'\tilde{u}_{1s}, Z_{2s}'\tilde{u}_{2s})'\). Then \(\hat{S}_{hh} = T^{-1} \sum_{s=1}^{T} \hat{h}_s'\hat{h}_s\). For \(i = 1, 2\), the \((i,i)\) block of \(\hat{B}\) is \((T^{-1} \sum_{s=1}^{T} Z_{is}'X_{is})^{-1}\); the off-diagonal blocks are zero. (ii) As suggested in comment 6 in Section 4, \(S_{ff}\) and \(\tilde{F}\) were computed using the out of sample residuals. Let \(\hat{f}_{t+1} = (\hat{u}_{1t+1}, \hat{u}_{2t+1})'\). Then \(\hat{S}_{ff} = P^{-1} \sum_{t=R}^{T} (\hat{f}_{t+1} - \hat{f}_T)(\hat{f}_{t+1} - \hat{f}_T)'\), and \(\hat{F} = (0, -2P^{-1} \sum_{t=R}^{T} w_{1t+1} \hat{u}_{1t+1}, 0, 0, 0, -2P^{-1} \sum_{t=R}^{T} w_{2t+1} \hat{u}_{2t+1})'\).

Let \(\alpha = (1 - 1)'\). The test statistic is \(P(\alpha'\tilde{f})^2/(\alpha'\hat{S}_{ff}\alpha) \sim \chi^2(1)\), where \(\hat{\Omega} = \hat{S}_{ff} + 2\Pi \tilde{F}\hat{V}_\beta \tilde{F}'\), \(\hat{V}_\beta = B'\hat{S}_{hh}B\). A test statistic that ignores uncertainty about \(\beta^*\) is \(P(\alpha'\tilde{f})^2/(\alpha'\hat{S}_{ff}\alpha)\); its seems of interest to evaluate such a statistic since it seems to be one used by Meese and Rogoff (1988) and suggested by Diebold and Mariano (1994). For such an evaluation to be empirically relevant, one would want an empirically plausible ratio of \(\alpha'FV_\beta F'\alpha\) (the term due to uncertainty about \(\beta^*\)) to \(\alpha'\hat{S}_{ff}\alpha\) (the term that would be present even if \(\beta^*\) were known): the test statistic that ignores \(\alpha'FV_\beta F'\alpha\) will behave arbitrarily poorly for an arbitrarily high ratio of \(\alpha'FV_\beta F'\alpha\) to \(\alpha'\hat{S}_{ff}\alpha\). Since predictive accuracy figures prominently in the exchange rate literature, I fit a convenient model (Meese (1986)). As described in a footnote, the resulting estimates are consistent with this experiment’s ratio of \(\alpha'FV_\beta F'\alpha\) to \(\alpha'\hat{S}_{ff}\alpha\).

In Table II, panel A presents results when a consistent estimator of \(\Omega\) is used, with \(\pi\) set to \(P/R\), while panel B presents results when the term due to uncertainty about \(\beta^*\) is ignored. In panel A, tests are slightly more poorly sized than in the previous experiment when \(P\) and \(R\) are both less than 50. The remaining entries are all between .049 and .075.

Panel B indicates that ignoring uncertainty about \(\beta^*\) can result in substantial overrejection, with nominal .05 tests having actual sizes larger than .50. As predicted by the theory, for given \(P\) the size distortions are smaller for larger \(R\).

---

4 I fit Meese’s (1986) model to monthly Deutschmark-dollar data 1974:3–1988:8, using a lag of Meese’s scalar measure of fundamentals as the instrument. The estimates of the scalars \(S_{ff}\) and \(FV_\beta F'\) yielded \(FV_\beta F' \approx S_{ff}\). With the DGP described above, it may be shown that \(\alpha'FV_\beta F'\alpha = 4\alpha'\hat{S}_{ff}\alpha\). So this experiment is, if anything, a little conservative about uncertainty about \(\beta^*\) (since \(4 < 5\)).
TABLE II
SIZE OF NOMINAL .05 TESTS ON EQUALITY OF MSPE, EXPERIMENT B

<table>
<thead>
<tr>
<th></th>
<th>A. π = P/R</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>25</td>
<td>50</td>
<td>100</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.022</td>
<td>0.035</td>
<td>0.057</td>
<td>0.071</td>
<td>0.075</td>
</tr>
<tr>
<td>50</td>
<td>0.027</td>
<td>0.036</td>
<td>0.058</td>
<td>0.064</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.049</td>
<td>0.056</td>
<td>0.063</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>B. π = 0</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>25</td>
<td>50</td>
<td>100</td>
<td>150</td>
<td>175</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.378</td>
<td>0.446</td>
<td>0.489</td>
<td>0.508</td>
<td>0.513</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.276</td>
<td>0.365</td>
<td>0.421</td>
<td>0.453</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.198</td>
<td>0.269</td>
<td>0.349</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: 1. As detailed in Section 5.2, the following was done 5000 times: a. 200 observations of the variables in a pair of linear regressions were generated. b. Each of these regressions was estimated 175 times by 2SLS, using observations 1–25, 1–26,…, and 1–199. The regressions have the same left-hand-side variable γt. The resulting structural coefficients and the realized values of the right-hand-side endogenous variables were used to predict yt one step ahead. c. For the values of R and P given above, and for each of the two competing models, a (P × 1) vector of prediction errors were formed using predictions dated \( R, R + 1, \ldots, R + P - 1 \). d. Equality of mean squared prediction error was tested, using the two vectors of sample prediction errors.

2. According to this paper's theory, the statistic in panel A is asymptotically \( \chi^2(1) \), that in panel B is not. The panel A statistic accounts for the fact that the predictions were made using estimated parameters; the panel B statistic is one that would be \( \chi^2(1) \) if the predictions were made using the population regression parameters rather than estimates. Each panel reports the fraction of the 5000 statistics that were greater than 3.84, which is the .05 critical value for a \( \chi^2(1) \) random variable.

(i.e., for a smaller ratio of \( P \) to \( R \); recall that as \( \pi = \lim(P/R) \to 0 \), the term in \( \Omega \) due to uncertainty about \( \beta^* \to 0 \).

One cautionary note: Recent literature suggests circumstances under which it may be difficult to obtain accurately sized in-sample tests (e.g., Newey and West (1994), Nelson and Startz (1991)). Such circumstances no doubt lead to poorly sized out-of-sample tests as well. In that sense, the simulation evidence presented here probably is unduly supportive of the asymptotic approximation.

6. CONCLUSION

Priorities for future research include: allowing for functions that are not differentiable, such as mean absolute prediction error and functions of the predicted sign of a left-hand-side variable; allowing for nonparametric estimators of regression parameters; application of the results in this paper to computationally convenient regression tests of predictive ability; and analysis of the power of tests applied to the moments of out-of-sample prediction errors, in a framework broad enough to include models that are overfit in-sample.

Dept. of Economics, University of Wisconsin, 1180 Observatory Dr., Madison, WI 53706-1393, U.S.A.

Manuscript received May, 1994; final revision received October, 1995.
APPENDIX

NOTATION: "\text{"{s}up}_{t\geq R\leq T}" means "\text{"{s}up}_{R\leq t\leq T}"; "\text{var}", "\text{cov}" denote variance and covariance; all limits are taken as the sample size \( T \) goes to infinity; summations of variables indexed by \( t \), \( t + \tau \), or \( t + j \) run from \( t = R \) to \( t = R + P - 1 = T \); summations of variables indexed by \( s \) run from \( s = 1 \) to \( s = t \).

LEMMA A1: (a) For \( 0 < a < 0.5 \), \( P^{-1/2} \sum t^{-1+a} \to 0 \); (b) \( P^{-1/2} \sum t^{-1/2} = O(1) \).

PROOF: (a) Choose \( a, 0 < a < 0.5 \). For \( \tau < \infty \), use \( P^{-1/2} \sum t^{-1+a} \leq P^{-1/2}(PR^{-1+a}) = P^{-1/2} \sum t^{-1+a} \to 0 \). For \( \tau = \infty \), use \( P^{-1/2} \sum t^{-1+a} \leq P^{-1/2} \int_{t=1}^{T} x^{-1+a} dx \leq a^{-1}P^{-1/2}(R+P-1)^a \to 0 \). For \( a = 0 \) the result follows since \( \sum t^{-1+a} \) is increasing in \( a \). (b) Follows by setting \( a = 0.5 \) in the proof of part (a).

LEMMA A2: Let \( x_t = [\text{vec}(f_{iB} - F') y_t, (f_t - Ef_t)' h_t]' \), \( \gamma_t = Ex_t \). Then \( (\sum_{t=1}^{T} |\gamma_t| \leq \infty \), (b) \( P^{-1/2} \sum t^{-1/2} \sum_{|\gamma_t| > 0} \gamma_t \to 0 \).

PROOF: (a) In light of Ibragimov and Linnik (1971, p. 307), (a) follows from Assumption 3. (b) follows from part (a) and Lemma A1(a).

LEMMA A3: For \( 0 < a < 0.5 \), (a) \( \text{sup}_t \|t^a H(t)\| \leq 0 \); (b) \( \text{sup}_t \|t^a (\hat{H}_t - \beta^*)\| \leq 0 \).

PROOF: (a) In this proof only, \( [\cdot] \) denotes "integer part of," \( q = 1 \), and \( k_t \) is an unimportant constant. Let \( \alpha_t \) be the mixing coefficients. Note that by Assumption 3, \( \alpha_t = \text{size of -3d/(d-1)} \) \( \Rightarrow \alpha_t = \text{size of -2d/(d-1)} \) \( \Rightarrow h_t \) is a mixingale with Hall and Heyde’s (1980, p. 19) symbols \( c_n, \psi_n \) and \( \phi_n \) defined as \( c_n = \text{max} \{1, \text{sup}_{t, 0} E|h_t|^2 \} \), \( \psi_n = \text{max} \{1, \text{sup}_{t, 0} E|h_t|^2 \} \), \( \phi_n = \text{max} \{1, \text{sup}_{t, 0} E|h_t|^2 \} \) (McLeish 1975, p. 837). Thus, \( \psi_n = k_t(|\gamma|/2)^{-3d/(d-1)} \) \( \Rightarrow \psi_n = k_t(|\gamma|/2)^{-3d/(d-1)} \). Since, finally, \( \sum_{t=1}^{T} \|t^{-1+a}\|^2 \leq \infty \) for \( 0 < a < 0.5 \), the result follows from Hall and Heyde (1980, Theorem 2.21). (b) By Assumption 2, \( \text{sup}_t \|B(t) - B\| \leq 0 \).

LEMMA A4: (a) \( P^{-1/2} \sum (f_{i\tau,\tau} - F)BH(t) = \alpha_p(1) \); (b) \( P^{-1/2} \sum (f_{i\tau,\tau} - F)B(t) - B)H(t) = \alpha_p(1) \); (c) \( P^{-1/2} \sum \|B(t) - B\|H(t) = \alpha_p(1) \).

PROOF: (a) Let \( v_t = (f_{i\tau,\tau} - F) \), redefine \( Bh_t \) as \( h_t \) and \( \gamma_t = Ex_t \). We have \( \|E^{1/2} \sum v_t H(t)\| \leq \|E^{1/2} \sum \gamma_t H(t)\| \leq \|P^{-1/2} \sum _{t=1}^{T} \gamma_t H(t)\| \to 0 \) by Lemma A1(a). Assumption 3(a) bounds the fourth moments of \( \text{vec}(\gamma_t)'h_t \). To \( \text{vec}(\gamma_t)'h_t \), in such a way that the \( \text{im} \text{var}(\sum_{t=1}^{T} \gamma_t H(t)) = 0 \) (West 1994) and the result follows from Chebyshev’s inequality. (b) Let \( v_t = (f_{i\tau,\tau} - F) \) and for simplicity assume \( \tau = k = -1 \). We have \( \|E^{1/2} \sum v_t (B(t) - B)H(t)\| \leq \|E^{1/2} \sum v_t H(t)\| \) and since \( \|E^{1/2} \sum v_t (B(t) - B)\| \leq 0 \) by Assumption 2 it suffices to show \( P^{-1/2} \sum v_t H(t) = \alpha_p(1) \). From Assumption 3, \( E[\gamma_t H(t)] \leq \|t^{-1/2} \| \) for a constant \( c \) that does not depend on \( t \); in the fourth order stationary case, for example, \( c = E(\gamma_t) \sum_{t=1}^{T} \gamma_t H(t) \leq P^{1/2} \sum (Eh_t h_t)^2 + \sum_{t=1}^{T} \gamma_t H(t) \leq \|cP^{-1/2} \sum t^{-1/2} = O(1) \) by Lemma A1(b) and the result follows from Markov’s inequality. (c) By logic such as that in the proof of Lemma A4(b), \( P^{-1/2} \sum H(t) = \alpha_p(1) \). Then \( \|E^{1/2} \sum B(t) - B\|H(t) \leq q \|E^{1/2} \sum B(t) - B\|P^{-1/2} \sum \|H(t)\| \leq 0 \) by Assumption 2.
PROOF OF EQUATION (4.1): A second order mean value expansion of \(f_{x+}(\hat{\beta}_t)\) around \(\beta^*\) for \(t = R, \ldots, T\) yields
\[
f_{x+}(\hat{\beta}_t) = f_{x+}(\hat{\beta}_t) + f_{x+}(\beta^*) + w_{t+\tau},
\]
where the \(i\)th element of \(w_{t+\tau}\) is \(w_{t+\tau} = \mathcal{S}(\hat{\beta}_t - \beta^*) \mathcal{S}(\hat{\beta}_t - \beta^*)\). The \(i\)th element of \(w_{t+\tau}\) is \(w_{t+\tau} = \mathcal{S}(\hat{\beta}_t - \beta^*) \mathcal{S}(\hat{\beta}_t - \beta^*)\). Then \(P^{1/2}(\hat{\beta}_t - \beta^*) = P^{1/2}(\hat{\beta}_t - \beta^*) + P^{1/2} \mathcal{S}(\hat{\beta}_t - \beta^*)\) and \(P^{1/2}(\hat{\beta}_t - \beta^*)\) will follow if (a) \(P^{1/2} \mathcal{S}(\hat{\beta}_t - \beta^*) = P^{1/2} \mathcal{S}(\hat{\beta}_t - \beta^*)\) and (b) \(P^{1/2} \mathcal{S}(\hat{\beta}_t - \beta^*) = \mathcal{S}(\hat{\beta}_t - \beta^*)\).

PROOF OF (a): We have
\[
P^{-1/2} \sum f_{x+}(\hat{\beta}_t - \beta^*) = P^{-1/2} \sum f_{x+}(\hat{\beta}_t - \beta^*) B(t) H(t)
= FB \left[ P^{-1/2} \sum H(t) \right] + P^{-1/2} \sum (f_{x+}(\hat{\beta}_t - \beta^*) - F) B(t) H(t)
+ P^{-1/2} F \sum (B(t) - B) H(t)
+ P^{-1/2} \sum (f_{x+}(\hat{\beta}_t - \beta^*) - F) (B(t) - B) H(t).
\]

The second term after the equality is \(o_p(1)\) by Lemma A4(a), the third by Lemma A4(c), the fourth by Lemma A4(b).

PROOF OF (b): Choose \(a, 0 < a < 0.5\) and define \(m_t\) as in Assumption 1. Then for the \(i\)th element of \(w_{t+\tau}\),
\[
\left| \sum w_{t+\tau} \right|
\leq 5 \sum_{t} \left| \mathcal{S}(\hat{\beta}_t - \beta^*) \right|^2 P^{-1/2} \sum \left| \mathcal{S}(\hat{\beta}_t - \beta^*) \right|^2 P^{1/2} \sum m_{t+\tau} P^{1/2} \leq 0,
\]
by Lemma A3(b), Lemma A1(a), Assumption 1, and Markov's inequality.

LEMMA A5: \(\lim E[ P^{-1} \sum H(t) \sum H(t)'] = 2[1 - \pi^{-1} \ln(1 + \pi)] S_{hh} = 2 I S_{hh} \).

PROOF: Set \(q = 1\) for notational simplicity. Let
\[
\gamma_t = \Gamma_{hh}(j), \quad a_{R,s} = (R + s)^{-1} + \cdots + (R + P - 1)^{-1} \quad \text{for} \quad 0 \leq s \leq P - 1,
\]
where the dependence of \(a_{R,s}\) on \(P\) is suppressed for notational convenience. L'Hôpital's rule indicates that for \(\pi = 0, 1 - \pi^{-1} \ln(1 + \pi) = 0\). To establish the lemma in this case, proceed as follows: Since
\[
\sum H(t) = a_{R,0}(h_1 + \cdots + h_R) + a_{R,1}h_{R+1} + \cdots + a_{R,P-1}h_{R+P-1},
\]

\[
\text{var}(P^{-1/2} \sum H(t)) \leq \sum_{j=1}^{\infty} \left| \gamma_t \right| P^{-1} \left( a_{R,0}^2 h_1^2 + a_{R,1}^2 + \cdots + a_{R,P-1}^2 \right);\]
since \(a_{R,j} \leq (P-j)/R\),
\[
P^{-1}\{Ra_{R,0}^2 + a_{R,1}^2 + \cdots + a_{R,P-1}^2\}
\leq \left(\frac{R}{P}\right)\left(\frac{P}{R}\right)^2 + (PR^2)^{-1}[1 + 2^2 + \cdots + (P-1)^2]
= \left(\frac{P}{R}\right) + (PR^2)^{-1}O(P^3)
= O(P/R) \rightarrow 0
\Rightarrow \text{var} \left( P^{-1/2} \sum H(t) \right) = 0.
\]

For \(\pi > 0\), write
\[
P^{-1/2} \sum H(t) = A_1 + A_2,
A_1 = P^{-1/2}a_{R,0}(h_1 + \cdots + h_R),
A_2 = P^{-1/2}[a_{R,1}h_{R+1} + \cdots + a_{R,P-1}h_{R-P-1}]
\Rightarrow \text{var} \left( P^{-1/2} \sum H(t) \right) = \text{var}(A_1) + \text{var}(A_2) + 2 \text{cov}(A_1, A_2).
\]

It will be shown that when \(\pi > 0\),
\[
(A-1a) \quad \lim \text{var}(A_1) = \pi^{-1}\ln^2(1 + \pi) \sum_{j=1}^{\infty} \gamma_j,
\]
\[
(A-1b) \quad \lim \text{var}(A_2) = \{2[1 - \pi^{-1}\ln(1 + \pi)] - \pi^{-1}\ln^2(1 + \pi) \} \sum_{j=1}^{\infty} \gamma_j,
\]
\[
(A-1c) \quad \lim \text{cov}(A_1, A_2) = 0.
\]

From the definition of \(A_1\), \(\text{var}(A_1) = P^{-1}a_{R,0}^2 \sum_{j=1}^{R-1} \gamma_j = (R/P)a_{R,0}^2 \sum_{j=1}^{R-1} \gamma_j = P^{-1}a_{R,0}^2 \sum_{j=1}^{R-1} \gamma_j\). In light of Lemma A2(b), (A-1a) will follow if \((R/P)a_{R,0}^2 \rightarrow \pi^{-1}\ln^2(1 + \pi)\).

We have
\[
\ln[(R + P - 1)/R] = \int_0^{P-1} (R + x)^{-1} dx \leq a_{R,0} \leq \int_{-1}^{P-1} (R + x)^{-1} dx
\]
\[
= \ln[(P+1)/R] \leq (R/P) \ln^2((R+P-1)/R) \leq (R/P) \ln^2((R+P-1)/(R-1))
\]
\[
= (R/P) a_{R,0}^2 \rightarrow \pi^{-1}\ln^2(1 + \pi),
\]
where the result for \(\pi = \infty\) follows since L'Hôpital's rule indicates that \(x^{-1}\ln^2(1+x) \rightarrow 0\) as \(x \rightarrow \infty\).

Now consider (A-1b). Let
\[
d_j = a_{R,1}a_{R,j+1} + \cdots + a_{R,P-j-1}a_{R,P-1} \quad \text{for} \quad 0 \leq j \leq P - 2;
\]
\[
d_j = a_{R,j} \quad \text{for} \quad P - 2 < j < 0;
\]
\[
\Rightarrow \text{var}(A_2) = P^{-1} \sum_{j=-P+2}^{P-2} d_j \gamma_j = P^{-1}d_0 \sum_{j=-P+2}^{P-2} \gamma_j + P^{-1} \sum_{j=-P+2}^{P-2} (d_j - d_0) \gamma_j,
\]
where the dependence of \(d_j\) on \(P\) and \(R\) is suppressed for notational simplicity. Using logic such as that in the proof of (A-1a), we have \(P^{-1}d_0 = P^{-1} \int_1^{P+1} \ln^2((R + P - 1)/(R + y))dy + o(1)\), where "\(o(1)\)" denotes a sequence whose limit is zero. Routine manipulations then yield
\[
P^{-1}d_0 = P^{-1}2(P - 2) - 2(R + 1)\ln[(R + P - 1)/(R + 1)]
- (R + 1)\ln^2((R + P - 1)/(R + 1)) + o(1)
\Rightarrow P^{-1}d_0 \rightarrow 2[1 - \pi^{-1}\ln(1 + \pi)] - \pi^{-1}\ln^2(1 + \pi)
\]
again using L'Hôpital's rule for \(\pi = \infty\).
INFERENCES ABOUT PREDICTIVE ABILITY

So (A.1b) will follow if \( P^{-1} \sum_{j=0}^{p-2} (d_j - d_0) \gamma_j \to 0 \). To show this, let \( m = \min((P/2) - 1, R) \).

Write \( P^{-1} \sum_{j \leq m} (d_j - d_0) \gamma_j = P^{-1} \sum_{j \leq m} (d_j - d_0) \gamma_j + P^{-1} \sum_{j > m} (d_j - d_0) \gamma_j \). By the Cauchy-Schwarz inequality, \( 0 \leq d_j \leq d_0 \Rightarrow |d_j - d_0| \leq d_0 \), so \( P^{-1} \sum_{j > m} (d_j - d_0) \gamma_j \to 0 \) since \( P^{-1} d_0 \) is bounded, \( m \to \infty \), and \( \sum_{j=m+1}^{\infty} |\gamma_j| < \infty \). To analyze \( P^{-1} \sum_{j \leq m} (d_j - d_0) \gamma_j \), note that for \( 0 \leq j \leq m \), we have

\[
d_0 - d_j = \frac{1}{2} \left[ (a_{R,j+1} - a_{R,j})^2 + \cdots + (a_{R,p-1} - a_{R,p-1-j})^2 \right] + \frac{1}{2} \left[ a_{R,1}^2 + \cdots + a_{R,j-1}^2 + a_{R,j}^2 + \cdots + a_{R,p-1}^2 \right] \leq \frac{1}{2} (P-j-1) [j/(R+1)]^2 + \sum_{j=1}^{j-1} \gamma_j^2 \leq \frac{1}{2} P [j/(R+1)]^2 + \frac{1}{2} \gamma_j^2 \leq \frac{1}{2} S \gamma_j^2,
\]

and \( P^{-1} \sum_{j \leq m} (d_j - d_0) \gamma_j \to 0 \) now follows from Lemmas A2(a) and A2(b).

Finally, consider (A.1c). Since \( a_{R,j} \leq a_{R,0} \), a direct calculation yields \( \text{cov}(A_1, A_2) \leq \frac{1}{2} \gamma_j^2 \). In particular, \( \text{cov}(A_1, A_2) \leq \frac{1}{2} \gamma_j^2 \to 0 \) by Lemma A2(b).

**Lemma A6:** lim \( E(P^{-1} \sum_{j=1}^{t} (f_{j+1} - E_{j+1}) \sum_{j=1}^{t} H(j)^t) = [1 - \pi^{-1} \ln(1 + \pi)] \sum_{j=1}^{t} \mu_{j+1} \mu_j = \pi S_{j+1} \).

**Proof:** Follows from an argument similar to that for Lemma A5 (West (1994)).

**Proof of Lemma 4.1:** That the limiting variance of \( P^{-1/2} \sum_{j=1}^{t} (f_{j+1} - E_{j+1}) \sum_{j=1}^{t} H(j)^t \) is as indicated follows from Lemmas A5 and A6. That \( V(\pi) \) is of full rank when \( \pi \neq 0 \) and \( S \) is p.d. is straightforward to establish. Asymptotic normality follows from McLeish (1977) as shown in Theorem 3.1 of Wooldridge and White (1989).

**Proof of Theorem 4.1:** Follows from equation (4.1) and Lemma 4.1.

**References**


