McCarthy, M. D., 'A note on the forecasting properties of two stage least squares restricted reduced forms—the finite sample case', *International Economic Review*, **13** (1972), 757–61.
Sargan, J. D., 'The existence of the moments of estimated reduced form coefficients', Unpublished paper, London School of Economics, 1976. Printed in Maasoumi, E. (ed.), *Contributions to Econometrics: John Denis Sargan*, Vol. 2, Cambridge: Cambridge University Press, 1988, pp. 133–57.

On the Limitations of Comparing Mean Square Forecast Errors: Comment

KENNETH D. WEST
University of Wisconsin, USA

In this stimulating paper, Clements and Hendry (CH) point out that rankings based on estimated MSFEs may be intransitive across linear models, with one model producing the lowest estimated MSFE for (say) log GNP, another for GNP growth (the difference of log GNP). They also suggest a different, and, in principle, less awkward, criterion, denoted GFESM, which, for a given horizon, is guaranteed to produce transitive rankings for any linear transformation of variables to be forecast, such as from logs to log differences.

CH's analysis of possible problems with use of MSFE is one that every researcher using that criterion should keep in mind, and their alternative criterion is a useful addition to the kit of tools we can use to evaluate models. My discussion will therefore be a series of loosely connected points related to the issues that CH have raised, rather than a critique of CH's basic message.

Let me begin by noting that, superficial appearances perhaps to the contrary, CH's discussion is perfectly consistent with standard results on optimal linear predictors under mean squared error loss. Let $\mathbf{x}_t$ be stationary with independent innovations, and consider forecasts that are linear functions of the past history of $\mathbf{x}_t$. Then if one predicts $\mathbf{x}_{t+h}$ in standard fashion from $\mathbf{x}_t$'s ARMA representation, using the population ARMA parameters, such a predictor will be optimal according to the criterion that CH call MSFEM (Granger and Newbold, 1977, p. 228) and therefore by what they call TMSFE as well.

A quick skim suggests that Granger and Newbold (1977) do not explicitly discuss forecasts when $\mathbf{x}_t$ has been linearly transformed or filtered. But linearity being what it is, the predictor that is optimal by MSFEM and TMSFE is the one obtained from the obvious linear transform or filter of $\mathbf{x}_t$'s ARMA representation. Suppose, for example, that $\mathbf{x}_t \equiv x_t$ follows a scalar AR(1) as in equations (7) to (13), $x_t = \rho x_{t-1} + u_t$. Then when forecasting Δx_{t+h} ($h > 0$), the linear function of past x_t's with the smallest MSFE is $(\rho^h - \rho^{h-1})x_t$. Or suppose that we have the bivariate system described in equations (14)–(16). Let $\mathbf{x}_t \equiv (o_t : m_t)'$ have as its infinite MA representation $\Sigma_{j=0}^{\infty}\psi_j u_{t-j}$. Let $\mathbf{x}_t^* = \mathbf{M}\mathbf{x}_t$ for some non-singular (2×2) matrix $\mathbf{M}$, perhaps $\mathbf{M}$ as defined above equation (14). Then the predictor of $\mathbf{x}_{t+h}^*$ that is optimal by MSFEM and TMSFE is $\Sigma_{j=0}^{\infty}\mathbf{M}\psi_{j+h}u_{t-j}$.

More generally, for stationary $\mathbf{x}_t$, use of the population ARMA parameters results in forecasts that are optimal by these criteria, for levels or first or nth differences, for this or that or the other linear transformation of $\mathbf{x}_t$, for any and all horizons. CH point out, however, that, in practice, the MSFEM and TMSFE criteria may inconsistently recommend different models

for different linear transformations of $\mathbf{x}_t$. Because of the effects of estimation uncertainty, this might happen even if one of the models under consideration is a 'correctly' parameterized ARMA model for $\mathbf{x}_t$.

What, then, is one to do? Clements and Hendry wisely refrain from making absolute and unqualified recommendations, instead concluding that 'more attention should be paid to ... invariant criteria'. Let me complement their discussion with two points on the use of MSFEs and other criteria in forecast evaluation.

First, on this side of the Atlantic at least, it is conventional wisdom that what measure of model performance is appropriate depends on how one plans to use the analysis. My own view is that this applies even when one is forecasting from a given set of data. For example, I have considered model evaluation when one is choosing among models of conditional variances because one is a risk-averse utility maximizer and wants to use an estimated conditional variance to make an asset allocation decision (West *et al.*, 1993). It is natural (at least to me) to then measure the quality of a model by how much expected utility the model yields. It may be shown that this may be estimated by a certain asymmetric function of the error in forecasting the second-moment matrix of the vector of asset returns (e.g. the squared asset return, in the univariate case).

The point is that there is a natural measure of forecast quality, and no ambiguity about levels versus differences versus anything else. CH indicate that this is not the case in Engle and Yoo (1987), in which the Monte Carlo study was not connected to any specific empirical application. As somewhat of an outsider to the literature on forecast evaluation, I am not sure how often there is a natural measure, in those studies in which the investigator has a specific application in mind. But *if* there is a natural measure in a particular context, it perhaps is not of overwhelming importance that other measures that might be natural in other contexts might give different rankings.

My second and final point is that when use of MSFE brings conflicting results, it would seem to me to generally be of interest to ask why. Among the possibilities are that for the sample size and data available one model is good for one purpose (say, long horizon forecasts), a second for another (say, short horizon forecasts). A second possibility is that all models under consideration are roughly equivalent: if we could replicate the forecasting experiment over and over, keeping the size of the sample used for estimation fixed, each of the MSFEs for a set of horizons and linear transformations would, on average, be quite similar across models; since we observe only one experiment, it is a largely a matter of chance which model comes out best for a given horizon and linear transformation. A third possibility is that if we could replicate the experiment many times, it is very likely that one model would be estimated to be best for all horizons and transformations, but, by chance, this model did not fare particularly well in the one experiment actually observed.

No doubt there are other possibilities. What I hope CH will do is provide additional simulation evidence on (1) the likelihood, under various scenarios (such as those discussed in the previous paragraph), of MSFE criteria giving conflicting rankings for different horizons and linear transformations and (2) the usefulness and desirability of GFESM or other invariant criteria, under the various scenarios. I look forward to their research on these and related questions.

ADDITIONAL REFERENCE

West, K. D., Edison, H. J. and Dongchul Cho, 'A utility based analysis of some models of exchange rate volatility', forthcoming in the *Journal of International Economics*, 1993.