Correspondence On the Selection of Error Measures for Comparisons Among Forecasting Methods

J. Scott Armstrong University of Pennsylvania, Philadelphia, PA, USA
Robert Fildes The Management School, Lancaster University, UK

Reprinted with permission from *Journal of Forecasting, Vo. 14, 67-71 (1995)*

ABSTRACT

Clements and Hendry (1993) proposed the Generalized Forecast Error Second Moment (GFESM) as an improvement to the Mean Square Error in comparing forecasting performance across data series. They based their conclusion on the fact that rankings based on GFESM remain unaltered if the series are linearly transformed. In this paper, we argue that this evaluation ignores other important criteria. Also, their conclusions were illustrated by a simulation study whose relationship to real data was not obvious. Thirdly, prior empirical studies show that the mean square error is an inappropriate measure to serve as a basis for comparison. This undermines the claims made for the GFESM.

KEY WORDS Accuracy Forecast evaluation Loss functions

Introduction

Presumably, one engages in theoretical manipulations and simulations in order to make generalizations about the real world. The purpose of Clements and Hendry (1993) (henceforth referred to as C&H) was to make generalizations about the best metric for determining which of a set of forecasting methods is expected to be most accurate, They concluded that one should use the Generalized Forecast Error Second Moment (GFESM) rather than the Mean Square Forecast Error (which we refer to as the MSE). The heart of their argument was that whatever error measure is used, it should be invariant to scale-preserving linear transformations.

We believe that their conclusion lacks external validity. We discuss three major reasons. First, invariance of rankings to transformations is only one of many criteria that are helpful for examining forecast accuracy. Second, the simulated data were not shown to provide a good representation of real data. Third, empirical studies have shown that mean square errors are inappropriate for the comparison of forecasting methods across different data series (Armstrong and Collopy, 1992; Fildes, 1992). Thus, the MSE is not a good benchmark.

Some of these issues are touched on by the discussants of C&H. We highlight what we consider to be the crucial issues, and refer to some relevant empirical studies that were overlooked by C&H and their commentators.

Criteria For Adopting Error Measures

A forecasting method generates a multivariate error distribution, conditioned by the various forecasting lead times under consideration and the time origin of the forecasts (Murphy and Winkler, 1992). The purpose of an error measure is to provide an informative and clear summary of the distribution.

Many commentators, including those who contributed to the discussion of C&H, believe that if a well-specified loss function exists, it should be used for the evaluation. Zellner (1986), Diebold (1993), and Granger (1993) go further, arguing that it should also be used in the model's estimation. However, Fildes and Makridakis (1988) point out that the available evidence is in conflict with this proposition. A well-specified loss function, while desirable, cannot be regarded as sufficient. In particular, operational assumptions need to be made about the error distribution. Loss functions for managers are typically in dollars and depend on the forecast errors in a complex way.[1]

Even if the loss function were known, typically it would be difficult to estimate. The distribution is usually unknown and it rarely conforms to the standard assumptions such as normality (Makridakis et al., 1987; Fildes, 1992; Armstrong and Collopy, 1994). Furthermore, other attributes of an error measure are important: error measures should be reliable (so that they can be used with small samples), resistant to outliers, and comprehensible to decision makers. The error measure should also provide a summary of the forecast error distribution for different lead times.

However, a loss function is rarely available to researchers. Nor is it common for forecasters in organizations to be clear as to their client's loss function. Using the C&H analogy, we doubt that the UK Chancellor of the Exchequer focuses solely on inflation despite the rhetoric, nor would we think his preferences are consistent over time.

Perhaps the ranking of forecasting methods should also be invariant to all scale-preserving transformations in the data, but this is not clear either to us or to the discussants.[2] For one thing, researchers who have been involved in studies comparing different measures have typically concluded that it is best to use dimensionless error measures, that is, those invariant to scalar transformations. At the very least then, the Mean Square Percentage Error would be preferred to the Mean Square Error.

Usually no single measure will capture the necessary complexity, largely because of lead-time effects. We believe that most users of forecasts are able to examine more that one error measure. Computer technology is making such comparisons easier. For example, statistical graphics now enable rich representations of the components of the error distribution. Wallis's (1993) graphic design is one example of such ingenuity. Software designers now include a variety of summary statistics and the issue should be which error measure or measures are appropriate for a given situation.

Simulated Data

C&H illustrated their approach using simulated data. They showed that the ranking of different estimation methods based on Mean Square Error depended on the representation of the data generation process adopted in the simulation, while that based on C&H's GFESM remained the same. The commentators (e.g., Engle, 1993) found this neither surprising nor undesirable. As Engle points out, the proposed error measure is based on one-step-ahead errors, while the divergence in estimator performance (the focus of the illustration) arose in multi-step-ahead predictions. Thus, the rankings based on a one-step-ahead measure did not change in contrast to a ranking based on a multi-step-ahead measure, illustrating that different loss functions may be needed to match the disparate objectives of a simulation or forecasting comparison.

In simulation studies, the burden of proof should be on the researcher to demonstrate that the conclusions can be generalized to the analysis of real data. C&H do not do this, and therefore, their conclusions as to the merits of their chosen error measure are not well supported.

Mean Square Errors

Does MSE have a role to play in forecasting comparisons, whether used in its original form or C&H's variant? While C&H recognized some of the problems with using MSE as a summary measure across series, they might have justified their interest by the high rating by forecasting practitioners and researchers in the survey by Carbone and Armstrong (1982). However, research since 1982 has cast considerable doubt on the usefulness of

[1] Some work has been done to try to relate such loss functions to error measures (e.g. Gardner, 1990; Price and Sharp, 1986).

[2] Schmidt (1993) points out that C&:H's proposed measure, GFESM, is invariant to scalar transformations, unlike MSE.

MSE-based measures. The growing consensus among researchers who have been making comparisons among forecasting methods is that the MSE should not be used. For example, Newbold (1983) criticized its use in the M-competition (Makridakis et al., 1982). Thompson (1990) studied this issue and concluded that the MSE was not appropriate. He also proposed a variation on the MSE, the log mean squared error ratio (LRM), that would be appropriate for making comparisons across series. The LMR takes the log of the ratio calculated by dividing the proposed model's MSE by the MSE of a benchmark model. A similar variation of the MSE had been used by Fildes and Makridakis (1988), whereby the MSE for a benchmark model was subtracted from that for a proposed model. In each study, the solution was to compare the MSE with that of a benchmark model.

Not all researchers agree, though. Zellner (1986), in an evaluation of the performance of Bayesian forecasting, concluded that the Bayesian approach produced more accurate forecasts in the M-competition when measured by MSE. This was criticized by Fildes and Makridakis (1988), in particular because the MSE results for five series dominated the results from the study of the full 1001 series (Chatfield, 1988). Such results have led researchers to recommend measures that control for scale.

Fortunately, as noted, empirical studies have been done using actual data to evaluate the usefulness of alternative measures, including MSE (Armstrong and Collopy, 1992; Fildes, 1992). These studies concluded that MSE should not be used for making comparisons among forecasting methods because it is unreliable. It is also sensitive to non-normal data contaminated by outliers, and such data are common. It is also difficult for users to understand. The alternative proposed by C&H seems to be subject to the same problems.

Conclusions

The available empirical evidence seems clear. Neither the MSE not the GFESM have a place in comparing the relative accuracy of forecasting methods across data sets. The issue had been addressed previously with the conclusion that relative error measures should be used. Empirical research has supported that position. The C&H simulation based on limited criteria does little to alter the weight of the evidence.

Acknowledgements

The US Navy Personnel R&D Center in San Diego, California, provided partial support for this research.

References

Armstrong, J. S. and Collopy, F., "Error measures for generalizing about forecasting methods: Empirical comparisons," *International Journal of Forecasting*, 8 (1992), 69-80.

Armstrong, J. S. and Collopy, F., "Asymmetry of forecast errors: An empirical analysis of its causes" Working paper, Marketing Department, The Wharton School, University of Pennsylvania, 1994.

Carbone, R. and Armstrong, J. S., "Evaluation of extrapolative forecasting methods: Results of a survey of academicians and practitioners," *Journal of Forecasting*, 1 (1982), 215-17.

Chatfleld, C., "Apples, oranges and mean square error," *International Journal of Forecasting*, 4 (1988), 515-18.

Clements, M. P. and Hendry, D. F., "On the limitations of comparing mean square forecast errors," *Journal of Forecasting*, 12 (1993), 617-37.

Diebold, F. X., "On the limitations of comparing mean square forecast errors: Comment," *Journal of Forecasting*, 12 (1993), 641-2.

Engle, R. F., "On the limitations of comparing mean square forecast errors: Comment," *Journal of Forecasting*, 12 (1993), 642-4.

Fildes, R., “The evaluation of extrapolative forecasting methods,” *International Journal of Forecasting*, 8 (1992), 88-98.

Fildes, R. and Makridakis, S., “Forecasting and loss functions,” *International Journal of Forecasting*, 4 (1988), 545-50.

Gardner, E., “Evaluating forecast performance in an inventory control system,” *Management Science*, 36 (1990), 490-99.

Granger, C. W. J., “On the limitations of comparing mean square forecast errors: Comment,” *Journal of Forecasting*, 12 (1993), 651-2.

Makridakis, S. et al., “The accuracy of extrapolation (time series) methods: Results of a forecasting competition,” *Journal of Forecasting*, 1 (1982), 111-53.

Makridakis, S. et al., “Confidence intervals: An empirical investigation of the series in the M-competition,” *International Journal of Forecasting*, 3 (1987), 489-508.

Murphy, A. H. and Winkler, R. L., “Diagnostic verification of probability forecasts,” *International Journal of Forecasting*, 7 (1992), 435-55.

Newbold, P., “The competition to end all competitions,” *Journal of Forecasting*, 2 (19833, 276-9.

Price, D. H. R and Sharp, J. A., “A comparisons of different univariate forecasting methods in a model of capacity acquisition in UK electricity supply,” *International Journal of Forecasting*, 2 (1986), 333-48.

Schmidt, P., “On the limitations of comparing mean square forecast errors: Comment,” *Journal of Forecasting*, 12 (1993), 660-62.

Thompson, P. A., “An MSE statistic for comparing forecast accuracy across series,” *International Journal of Forecasting*, 6 (1990), 219-27.

Wallis, K. F., “On the limitations of comparing mean square forecast errors: Comment,” *Journal of Forecasting*, 12 (1993), 663-6.

Zellner, A., “A tale of forecasting 1001 series: The Bayesian knight strikes again,” *International Journal of Forecasting*, 2 (1986), 491-4.