7 Evaluating predictive accuracy

The question "How good is a forecast?" comprises two separate aspects: firstly, measuring predictive accuracy *per se*; secondly, comparing various forecasting models. For example, if a variable is almost unpredictable, all forecasts are likely to be poor. Yet, a forecaster may still look for the best forecast among the poor ones.

The most commonly reported measures of predictive accuracy are

1. mean squared prediction errors or a variant of them;
2. mean absolute prediction errors;
3. percentage measures, such as the *mean absolute percentage error* (MAPE);
4. Theil coefficients;
5. significance measures, such as the DIEBOLD-MARIANO test statistic.

Additionally, some researchers, particularly in applied economics, use or suggest *qualitative* accuracy measures. CHATFIELD focuses on such a measure under the name of 'Percent Better'. We should scrutinize each of these suggestions in turn.

7.1 Mean squared prediction errors

In the notation of CHATFIELD's textbook, the *prediction mean square error* PMSE is defined as

$$PMSE = m^{-1} \sum_{t=N-m+1}^{N} \left(x_t - \hat{x}_{t-1}(1) \right)^2 .$$

In this form, PMSE evaluates out-of-sample one-step errors. N observations are available, and the last m observations are used for evaluation. The forecast $\hat{x}_{t-1}(1)$ is meant to be calculated on the basis of the observations $x_1, \ldots, x_{t-1}$ only, including parameter estimates. The formula is easily modified for h–step errors with $h > 1$.

The PMSE formula does not specify m. A general recommendation is to keep m/N small. For large m/N, the forecasts with small t are based on very short samples and thus fail to be reliable indicators of absolute or

relative accuracy. An obvious modification would be to replace PMSE by a weighted MSE, with the weights increasing in t. Like the original PMSE, this suggestion emphasizes the possibility that the true model may change over time. Then, the approximation by the prediction model toward the end of the sample is potentially more important for forecasts beyond N than the approximation in the earlier portion.

As $m \to \infty$, the PMSE should converge to a variance. Depending on the properties of the DGP, if such a one is assumed to exist, that variance should be close to—though slightly larger than—the variance of the theoretical prediction error $\mathrm{E}\left(x_t - \mathrm{E}\left(x_t|\mathcal{I}_{t-1}\right)\right)^2$, where $\mathcal{I}_{t-1}$ is an information set that contains the history of the series x. That theoretical variance serves as the benchmark for the construction of many statistical procedures, including least-squares estimation and AIC, which justifies the widespread usage of the PMSE.

The most common critiques of the PMSE are:

1. quadratic loss may not correspond to the forecaster's loss function;
2. the PMSE depends on scales;
3. the PMSE is vulnerable to outliers.

The idea of a forecaster's loss function is that forecast errors entail costs, in the sense that costs depend on

$$\mathrm{E}\, g\left(x_t - \hat{x}_{t-1}\left(1\right)\right)$$

for some function g. Some authors even consider generalizations of the expectation measure, as the costs may depend on time or on a nonlinear transformation of the vector of prediction errors, or on a more qualitative evaluation. The function should obey $g\left(0\right) = 0$ and $g\left(x\right) > g\left(y\right)$ for $x > y > 0$ and $x < y < 0$. Otherwise, g may be asymmetric and it may converge to a finite constant as its argument approaches infinity. The main problem with the loss-function approach is that the true loss or cost function is rarely known in practice. In some applications, even the existence of a cost function is uncertain, as the forecast may satisfy curiosity rather than serve as a basis for actual immediate decisions. Therefore, only simple loss function are considered usually, such as $g\left(x\right) = x^2$ and $g\left(x\right) = |x|$. The first choice yields the PMSE, the second one yields the mean absolute error, which is occasionally preferred due to its robustness toward outliers.

The scale dependence of PMSE is not a problem, as long as a specific variable x is in focus. If several variables are predicted using various procedures in each case, individual PMSEs should be weighted. A suggestion for weighting would be the sample variance of each series.

Often, instead of the PMSE, its square root is reported. While the PMSE corresponds to a measure of variance, the root MSE is a measure of standard deviation, which makes it somehow easier to interpret. The transformation does not change the ranking of models and predictions according to their accuracy, and it also does not remove any of the inherent problems of the PMSE.

7.2 Mean absolute prediction errors

In the same notation as above, the *mean absolute error* is defined as

$$MAE = m^{-1} \sum_{t=N-m+1}^{N} \left| x_t - \hat{x}_{t-1}(1) \right|.$$

As $m \to \infty$, the MAE should converge to $\mathrm{E}\left| x_t - \mathrm{E}\left(x | \mathcal{I}_{t-1} \right) \right|$ or a slightly larger value, assuming this value exists. For a Gaussian world, this moment is proportional to the standard deviation, with a fixed proportionality factor. Also for other distributions, the absolute moment will measure the dispersion of the forecast errors.

The MAE is based on the loss function $g(x) = |x|$, which is more sensitive to small deviations from 0 and much less sensitive to large deviations than the usual squared loss. Therefore, the MAE can be viewed as a 'robust' measure of predictive accuracy. The MAE tends to prefer forecasting procedures that produce occasional large forecast failures, while they are reasonably good on average. By contrast, the MSE tends to prefer forecasting procedures that avoid large forecast failures, even though they produce a less satisfactory fit otherwise.

Because the estimation procedures are usually based on least-squares criteria, an emphasis on the MAE may involve a slight logical inconsistency. The best class of models is then selected according to a criterion that is different from the one that selects among the different members of an individual model class.

7.3 Mean absolute percentage error

In the above notation, the *mean absolute percentage error* (MAPE) is defined by

$$MAPE = m^{-1} \sum_{t=N-m+1}^{N} \left| \frac{x_t - \hat{x}_{t-1}(1)}{x_t} \right|.$$

This definitions answers complaints by some researchers that traditional criteria, such as PMSE and MAE, depend on the scaling of the variable x, which may be inconvenient if the criteria are used for comparing predictive accuracy across different variables or different time ranges. Unfortunately, the MAPE achieves scale independence by a simple division by x_t. This entails a serious drawback.

Many economic variables, such as stock returns and most growth rates, vary around zero. Whenever $x_t = 0$, the contribution at time point t and therefore the MAPE are undefined. Even if x_t is only approximately zero, the relative contribution of time point t will be enormous. Usually, there is no justification for preferring a high precision for small values of x_t.

It is obvious that the MAPE is tuned to variables that live in an area that is separated from zero by common sense. Economic examples would be the main aggregates of national accounts, such as fixed investment and private consumption.

Under the name of 'rmse percent error', PINDYCK AND RUBINFELD consider the direct squared-loss counterpart to the MAPE

$$m^{-1} \sum_{t=N-m+1}^{N} \frac{(x_t - \hat{x}_{t-1}(1))^2}{{x_t}^2}.$$

This measure has properties that are similar to those of the MAPE, with whom it shares most of its problems. While PMSE is more often reported than MAE, MAPE appears to be more popular than the above suggestion.

In order to obtain a scale-free precision measure, it would be more appealing to consider measures such as

$$\frac{\sum_{t=N-m+1}^{N} (x_t - \hat{x}_{t-1}(1))^2}{\sum_{t=N-m+1}^{N} (x_t - \bar{x})^2}.$$

In this formula, the denominator measures the 'total' variation of x, while the numerator measures that part of the variation that has been accounted for by the prediction procedure. Thus, the measure is reminiscent of the regression R^2. THEIL's accuracy measures follow a similar idea.

7.4 Theil coefficients

The idea of Theil's coefficients was to evaluate a forecast against the background of a simple or primitive forecast. If a forecasting procedure is to be taken seriously, it should at least 'beat' the simple benchmark. Unfortunately, it is not always clear which benchmark to use. THEIL used mainly random-walk or no-change forecasts, while other researchers use autoregressive prediction or exponential smoothers instead.

The version of Theil's coefficient that has been implemented into the EViews software is defined as

$$U = \frac{\sqrt{\sum_{t=N-m+1}^{N} \left(x_t - \hat{x}_{t-1}\left(1\right)\right)^2}}{\sqrt{\sum_{t=N-m+1}^{N} x_t^2} + \sqrt{\sum_{t=N-m+1}^{N} \hat{x}_{t-1}\left(1\right)^2}}.$$

For a 'good' predictor, the numerator will be small compared to the denominator. For a 'bad' predictor, both will be of similar magnitude. Theil's measures have often been criticized in the literature. They tend to yield implausible results in the sense that a predictor that optimizes them may have undesirable properties and *vice versa*, at least under lab conditions. It is not so certain whether this is also true in practical applications.

7.5 Decomposing the mean squared error

According to the EViews manual and to PINDYCK&RUBINFELD, the mean squared forecast error can be decomposed as

$$m^{-1} \sum \left(x_t - \hat{x}_{t-1}\left(1\right)\right)^2 = \left(\sum \hat{x}_{t-1}\left(1\right) - \bar{x}\right)^2 + \left(s_{\hat{x}} - s_x\right)^2 + 2\left(1 - r_{x\hat{x}}\right) s_{\hat{x}} s_x.$$

Here, $s_{\hat{x}}$ and s_x are sample standard deviations of $\hat{x}$ and x, respectively, while $r_{x\hat{x}}$ is the sample correlation. Dividing the three parts by the total yields the bias proportion, the variance proportion, and the covariance proportion. The non-negative 'proportions' sum up to 1. According to the sources mentioned above, the *bias proportion* tells us how far the mean of the forecast is from the mean of the actual series. The *variance proportion* tells us how far the variation of the forecast is from the variation of the actual series. Finally, the *covariance proportion* measures the remaining unsystematic forecasting errors.

The idea is that, if the forecast is "good", the bias and variance proportions should be small so that most of the bias should be concentrated on the

covariance proportions. The informative value of the decomposition has not been accepted universally, however.

7.6 Diebold-Mariano statistics

The econometricians DIEBOLD and MARIANO were interested in a situation where a 'cheap' benchmark forecast is compared to a sophisticated forecast. A forecaster may prefer the cheap forecast up to a point where the sophisticated forecast shows its relative merits 'significantly'. It is uncertain whether this situation is common in applications and empirical projects. Usually, various forecasting methods are considered and an optimum method is then selected, while the cost of a forecast method play little role. A complicated forecasting method may even be selected if it only achieves a slight improvement on average. Some economic projects may even show an instinctive bias toward more sophisticated and costly methods, as these demonstrate the forecasting team's skills.

CHATFIELD ranks among the few statisticians who criticized the null hypothesis of these tests *expressis verbis*. That null hypothesis would be that the expected difference in squared error (or some other loss moment) is zero. It is doubtful whether this null hypothesis is of central interest to the typical forecaster.

DIEBOLD&MARIANO assume that the precision is basically measured by $\mathrm{E}\, g\left(x_t - \hat{x}_{t-1}\left(1\right)\right)$ and $\mathrm{E}\, g\left(x_t - \tilde{x}_{t-1}\left(1\right)\right)$ for two different forecasts $\tilde{x}$ and $\hat{x}$ and a loss function $g\left(.\right)$. Under the hypothesis that the difference is zero, it can be shown that the test statistic

$$S_1 = \frac{\bar{d}}{\sqrt{m^{-1} 2\pi \hat{f}_d\left(0\right)}}$$

converges to a standard normal distribution as $m \to \infty$. Here, $\bar{d}$ denotes the sample average of $d_t = g\left(x_t - \hat{x}_{t-1}\left(1\right)\right) - g\left(x_t - \tilde{x}_{t-1}\left(1\right)\right)$. The element $\hat{f}_d\left(0\right)$ is a scale factor defined as the spectral density estimate of d_t at the frequency 0. An operable definition for $\hat{f}_d\left(0\right)$ is

$$\hat{f}_d\left(0\right) = \left(2\pi\right)^{-1} \sum_{k=-m+1}^{m-1} w\left(k, m\right) \hat{\gamma}_d\left(k\right),$$

where $\hat{\gamma}_d\left(k\right)$ is the sample autocovariance at lag k and $w\left(.\right)$ is a kernel weight function that obeys certain consistency conditions, such as $w\left(k, m\right) \to 1$ for fixed k and $m \to \infty$.

7.7 Qualitative measures

Sometimes, macroeconomists maintain that they are less interested in the accuracy of a real growth forecast than in forecasting 'turning points' of the business cycle. It is uncertain whether such turning points exist outside of the official NBER chronology and economics textbooks. Similarly, a stock market analyst may be more interested in whether a specific security price is going to rise or to fall in the immediate future than in the numerical accuracy of the price prediction for the next day. Again, it is often uncertain whether the implied picture of longer sinusoidal swings in the security price with local maxima (peaks) and local minima (troughs) corresponds to reality.

It is difficult to construct an accuracy measure for this type of loss function, at least as long as the variable of concern x is quantitative. Some economists tend to 'code' and discretize some real-valued variables, such that 'x increases', 'x decreases', and 'x remains approximately constant' become the three events or 'states'. Then, one may count the occurrences of successes and failures. The 'better' procedure is the one that yields more successes or a larger success ratio. While the winning procedure may miss the exact value of x by far, its forecast for the 'sign' of x is reliable. Here, an inherent difficulty is the definition of 'approximate constancy'. Instead of success ratios, one may also summarize this type of evidence in coincidence matrices.

A different kind of success ratio is suggested by CHATFIELD who, in comparing two methods A and B, counts the cases when A is closer to the true value and when B is closer. A forecaster who predicts many values closely and misses a few ones by far, would attain a good 'Percent Better' ratio. In this sense, CHATFIELD's 'Percent Better' is a robust criterion and comparable to the MAE. Note, however, that it is scale-independent, unlike the MAE.

7.8 Forecast bias or mean error

The sample average of the prediction errors

$$m^{-1} \sum_{t=n-m+1}^{N} \left(x_t - \hat{x}_{t-1}\left(1\right)\right)$$

is also often reported for prediction experiments. It is not a real measure of accuracy, although it contains some important information. A forecast with a large and systematic forecast bias could be improved by some straightforward

adjustment. In this sense, systematic over-prediction or under-prediction points to an inefficiency, or otherwise to an asymmetric loss function.

References

[1] Anderson, T.W. (1951) 'Estimating linear restrictions on regression coefficients for multivariate normal distributions'. *Journal of the American Statistical Association* **85**, 813–823.

[2] Box, G.E.P., and G.M. Jenkins (1970) *Time Series Analysis, Forecasting, and Control*. Holden-Day.

[3] Brockwell, P.J., and R.A. Davis (1991) *Time Series: Theory and Methods*. 2nd edition, Springer-Verlag.

[4] Cartwright, N. (1995) 'Probabilities and experiments'. *Journal of Econometrics* **67**, 47–59.

[5] Chatfield, C. (2001) *Time-series Forecasting*. Chapman & Hall.

[6] Christoffersen, P.F., and Diebold, F.X. (1998) 'Cointegration and long-horizon forecasting', *Journal of Business & Economics Statistics* **16**, 450–458.

[7] Clements, M., and Hendry, D.F. (1998) *Forecasting economic time series*. Cambridge University Press.

[8] Dickey, D.A., and Fuller, W.A. (1979) 'Distribution of the estimators for autoregressive time series with a unit root' *Journal of the American Statistical Association* **74**, 427–431.

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

[10] Engle, R.F. (1992) 'Autoregressive conditional heteroskedasticity with estimates of variance of United Kingdom inflation' *Econometrica* **50**, 987–1007.

[11] Engle, R.F., Hendry, D.F., and Richard, J.-F. (1983) 'Exogeneity'. *Econometrica* **51**, 277–304.

[12] Engle, R.F., and Yoo, B.S. (1987) 'Forecasting and Testing in Cointegrated Systems', *Journal of Econometrics* **35**, 143–159.

[13] Franses, P.H. (1996) *Periodicity and Stochastic Trends in Economic Time Series*. Oxford University Press.

[14] Franses, P.H., and VanDijk, D. (2000) *Non-linear time series models in empirical finance*. Cambridge University Press.

[15] Gardner, E.S. Jr. (1985) 'Exponential Smoothing: The State of the Art' *Journal of Forecasting* **4**, 1–28.

[16] Gardner, E.S.Jr., and McKenzie, E. (1985) 'Forecasting trends in time series' *Management Science* **31**, 1237–1246.

[17] Granger, C.W.J. (1969) 'Investigating Causal Relations by Econometric Models and Cross-Spectral Methods' *Econometrica* **37**, 424–438.

[18] Granger, C.W.J. (1989) *Forecasting in Business and Economics*. Academic Press.

[19] Harvey, A.C. (1989) *Forecasting, Structural Time Series, and the Kalman Filter*. Cambridge University Press.

[20] Hylleberg, S., Engle, R.F., Granger, C.W.J., and Yoo, B.S. (1990) 'Seasonal integration and cointegration' *Journal of Econometrics* **44**, 215–238.

[21] Johansen, S. (1995) *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press.

[22] Pindyck, R.S. and Rubinfeld, D.L. (1991). *Econometric Models and Economic Forecasts*. 3rd edition, McGraw-Hill.

[23] Sims, C.A. (1980) 'Macroeconomics and reality'. *Econometrics* **48**, 1–48.

[24] Taylor, S. (1986) *Modelling Financial Time Series*. John Wiley & Sons.

[25] Tong, H. (1990) *Non-linear Time Series: A Dynamical Systems Approach*. Oxford University Press.