Testing Density Forecasts, With Applications to Risk Management

Jeremy Berkowitz
Graduate School of Management, University of California, Irvine, CA 92697-3125
(*jberkowitz@gsm.uci.edu*)

The forecast evaluation literature has traditionally focused on methods of assessing point forecasts. However, in the context of many models of financial risk, interest centers on more than just a single point of the forecast distribution. For example, value-at-risk models that are currently in extremely wide use form *interval forecasts*. Many other important financial calculations also involve estimates not summarized by a point forecast. Although some techniques are currently available for assessing interval and density forecasts, existing methods tend to display low power in sample sizes typically available. This article suggests a new approach to evaluating such forecasts. It requires evaluation of the entire forecast distribution, rather than a scalar or interval. The information content of forecast distributions combined with ex post realizations is enough to construct a powerful test even with sample sizes as small as 100.

KEY WORDS: Densities; Evaluation; Forecasting; Risk Management.

Although the forecast evaluation literature has traditionally focused on point forecasts, many models in economics and finance produce forecasts that cannot be easily summarized by a point forecast. For example, the widely used value-at-risk (VaR) approach to quantify portfolio risk delivers *interval forecasts* (see Jorion 1997 for a recent survey). These models are used to assess corporate risk exposures and have received the official imprimatur of central banks and other regulatory authorities. In 1997, a Market Risk Amendment to the Basle Accord permitted banks to use VaR estimates for setting bank capital requirements related to trading activity.

Many other important financial calculations involve estimates not summarized by a single point on the forecast density. For example, the Standard Portfolio Analysis of Risk (SPAN) system, first implemented by the Chicago Mercantile Exchange in 1988, has become a very widely used approach to calculate margin requirements for customers and clearinghouse members. SPAN is essentially a combination of stress tests performed on each underlying instrument in the relevant portfolio (see Artzner, Delbaen, Eber, and Heath 1999).

To evaluate the performance of such models and their forecasts, regulators and financial institutions must be able to compare the forecasts to subsequent outcomes. The most familiar such exercise is verifying that the model accurately delivers a given interval forecast. Early approaches to this problem were handled as if the upper end of the interval was a point forecast—the number of times the interval was exceeded is compared to the expected number.

However, Christoffersen (1998) emphasized that, being interval forecasts, there is more information in interval forecasts than a point. To see why, note that even if the model delivers the correct *average* coverage, it may not do so at every point in time. For example, in markets with persistent volatility, forecasts should be larger than average when volatility is above its long-run average and vice versa. Christoffersen proposed methods for evaluating interval forecasts. These methods go part of the way toward addressing the critique of Kupiec (1995), who argued that very large datasets are required to verify the accuracy of such models. Users typically set the VaR level deep in the tail of the distribution (the Accord stipulates a 99% level). Violations are thus expected to occur only once every 100 days. Simply counting the number of violations per year obviously uses very little of the data. Christoffersen suggested making greater use of the violations by noting that violations should occur 1% of the time and should not be bunched together—violations should be *conditionally* unpredictable. Nevertheless, interval evaluation methods remain quite data intensive since they only make use of whether or not a violation occurs (e.g., Lopez 1999).

If risk models are to be evaluated with small samples, a more general perspective on model performance is required. In particular, models can be more accurately tested by examining many percentiles implied by the model. To take this argument to its logical extreme, we might evaluate the entire forecast density. Forecasts at *every* percentile would then be compared to realized data. In this way, the time series information contained in realized profits and losses is augmented by the cross-sectional information in ex ante forecast *densities*. The additional information content is readily converted into testable propositions. Density forecast evaluation, however, is affected by the density's interior. Since the interior characterizes small day-to-day disturbances, it may be of substantially less concern to financial institutions, managers, and regulators than the tail behavior.

For this reason, attention has recently shifted to "exceedence" measures that account for the expected magnitude of large losses. In particular, Artzner et al. (1999) suggested the tail expected loss (EL) measure, $E(y_t|y_t < \bar{y}_t)$. These authors showed that this measure of risk fulfills a set of intuitively sensible axioms such as monotonicity and subadditivity while quantiles do not. Surprisingly, Artzner et al. also showed that scenario analyses such as SPAN can be cast as equivalent to an expected loss measure. Basak and Shapiro (1998) found, in the context of a stochastic equilibrium model, that risk

Journal of Business & Economic Statistics
October 2001, Vol. 19, No. 4