International Journal of Forecasting 19 (2003) 165–175

www.elsevier.com/locate/ijforecast

Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts

Kenneth F. Wallis*

Department of Economics, University of Warwick, Coventry CV4 7AL, UK

Abstract

This article reviews recently proposed likelihood ratio tests of goodness-of-fit and independence of interval forecasts. It recasts them in the framework of Pearson chi-squared statistics, and considers their extension to density forecasts. The use of the familiar framework of contingency tables increases the accessibility of these methods to users, and allows the incorporation of two recent developments, namely a more informative decomposition of the chi-squared goodness-of-fit statistic and the calculation of exact small-sample distributions. The tests are applied to two series of density forecasts of inflation, namely the US Survey of Professional Forecasters and the Bank of England fan charts. This first evaluation of the fan chart forecasts finds that, whereas the current-quarter forecasts are well-calibrated, this is less true of the one-year-ahead forecasts. The fan charts fan out too quickly and the excessive concern with the upside risks was not justified over the period considered.

Keywords: Forecast evaluation; Interval forecasts; Density forecasts; Likelihood ratio tests; Chi-squared tests; Exact inference; Bank of England inflation forecasts

JEL classification: C53; E37

1. Introduction

Interval forecasts and density forecasts are being increasingly used in practical real-time forecasting. An interval forecast of a variable specifies the probability that the future outcome will fall within a stated interval; this usually uses round numbers such as 50 or 90% and states the interval boundaries as the corresponding percentiles. A density forecast is an estimate of the complete probability distribution of the possible future values of the variable. As supplements to point forecasts they each provide a description of forecast uncertainty, whereas no information about this is available if only a point forecast is presented, a practice which is being increasingly criticized in macroeconomic forecasting. Density forecasts are more directly used in decision-making in the fields of finance and risk management. Tay and Wallis (2000) provide a survey of applications of density forecasting in macroeconomics and finance, and the examples in the present article come from the former field.

Evaluating the accuracy of interval and density forecasts is similarly receiving increasing attention. For interval forecasts the first question is whether the coverage is correct ex post, that is, whether the relative frequency with which outcomes are observed

*Tel.: +44-24-7652-3055; fax: +44-24-7652-3032.
E-mail address: k.f.wallis@warwick.ac.uk (K.F. Wallis).

0169-2070/02/$ – see front matter
doi:10.1016/S0169-2070(02)00009-2

to fall in their respective forecast intervals is equal to the announced probability. Christoffersen (1998) argues that this *unconditional* hypothesis is inadequate in a time-series context, and defines an efficient sequence of interval forecasts with respect to a given information set as one which has correct *conditional* coverage. He presents a likelihood ratio framework for conditional coverage testing, which combines a test of unconditional coverage with a test of independence. This supplementary hypothesis is directly analogous to the requirement of lack of autocorrelation of orders greater than or equal to the forecast lead time in the errors of a sequence of efficient point forecasts. It is implemented in a two-state (the outcome lies in the forecast interval or not) Markov chain, as a likelihood ratio test of the null hypothesis that successive observations are statistically independent, against the alternative hypothesis that the observations are from a first-order Markov chain.

For density forecasts, the first question again concerns goodness-of-fit. Two classical methods of testing goodness-of-fit—the likelihood ratio and Pearson chi-squared tests—proceed by dividing the range of the variable into k mutually exclusive classes and comparing the probabilities of outcomes falling in these classes given by the forecast densities with the observed relative frequencies. It is usually recommended to use classes with equal probabilities, so the class boundaries are quantiles; similarly, a standard way of reporting densities is in terms of their quantiles. For a sequence of density forecasts these change over time, of course. This approach reduces the density forecast to a k-interval forecast and sacrifices information, but the distinction between unconditional and conditional coverage extends to this case, and a likelihood ratio test of independence can be developed in a k-state Markov chain, generalizing Christoffersen's proposal above.

It is well known that the likelihood ratio tests and Pearson chi-squared tests for these problems are asymptotically equivalent; for general discussion and references to earlier literature see Stuart, Ord and Arnold (1999, chapter 25). In discussing this equivalence for the Markov chain tests they develop, Anderson and Goodman (1957) note that the chi-squared tests, which are of the form used in contingency tables, have the advantage that, 'for many users of these methods, their motivation and their application seem to be simpler', and this point of view prompts the present line of enquiry. In this article we accordingly explore the equivalent chi-squared tests for the hypotheses discussed above. The term chi-squared tests is used here and in the title of the article to refer to Pearson's chi-squared statistic, memorable to multitudes of students as the formula $\Sigma(O-E)^2/E$. Asymptotically it has the χ^2 distribution under the null hypothesis.

Two recent extensions to this framework are also considered. The first is the 'rearrangement' by Anderson (1994) of the chi-squared goodness-of-fit test to provide more information on the nature of departures from the null hypothesis, in respect of specific features of the empirical distribution such as its location, scale and skewness. Second, since our attention is focussed on applications in macroeconomics, where sample sizes are as yet small, we consider the exact finite-sample distribution of the chi-squared statistic. While the contingency table literature contains an extensive discussion of finite-sample behaviour, going back to Fisher's exact test—see Yates (1984) for a review of methods for 2×2 tables—it is only relatively recently that convenient computational routines have become easily available. The calculation of exact P-values based on the permutational distribution of the test statistic uses methods surveyed by Mehta and Patel (1998), implemented in StatXact-4 (Cytel Software Corp.).

Two datasets are used below. The first is the US Survey of Professional Forecasters' (SPF) density forecasts of inflation, previously analyzed by Diebold, Tay and Wallis (1999). The forecast densities are reported numerically, as histograms, and have no particular functional form. The fact that their skewness and kurtosis vary over time is well-established (Lahiri and Teigland, 1987), but there is no underlying model of this variation. The series of forecasts and outcomes are shown in Fig. 1, where plotted percentiles have been obtained by linear interpolation of the published histograms. The second example is the Bank of England Monetary Policy Committee's density forecasts of inflation, which date from the establishment of the Committee in mid-1997. A specific functional form is assumed, with three parameters that determine location, scale