

ON A TESTING PROCEDURE FOR MODEL SELECTION

J. Panaretos and S. Psarakis
 University of Patras, Greece

ABSTRACT

In this paper a forecasting model selection scheme is considered which amounts to testing the predictive behaviour of a model by adopting Xekalaki and Katti's (1984) idea of assigning to its performance a score for each of a series of time points. The score reflects how close to, or how far from, the predictive value the observed actual value is. A statistical test is proposed for comparing the forecasting performances of two models.

Selecting the best of two competing models

Consider two linear models A and B of the form

$$Y_t = X_t(M) \beta_M + e_t(M)$$

where Y_t is an 1×1 vector of observations on the dependent random variable, $X_t(M)$ is an $1 \times m_M$ matrix of known coefficients ($1 \geq m_M, |X_t(M)X_t(M)'| \neq 0$), β_M is an $m_M \times 1$ vector of regression coefficients, and $e_t(M)$ is an 1×1 vector of normal error random variables with $E(e_t(M)) = 0$ and $V(e_t(M)) = \sigma_M^2 I_t$ ($\sigma_M^2 < \infty$). Here I_t is the 1×1 identity matrix and M indexes the model (i.e. $M=A$ or $M=B$). Therefore, a prediction for the value of the dependent random variable for time $t+1$ will be given by the statistic $\hat{Y}_{t+1}^0(M) = X_{t+1}^0(M) \hat{\beta}_t(M)$, where $\hat{\beta}_t(M)$ is the least squares estimator of β_M at time t , $X_{t+1}^0(M)$ is a $1 \times m_M$ vector at time $t+1$. Let Y_{t+1}^0 be the observed value of the dependent random variable at time $t+1$. Then

$$\hat{Y}_{t+1}^0(M) - Y_{t+1}^0 = e_{t+1}^*(M) \text{ will follow the } N(0, \sigma_M^2) \quad (1)$$

and so $|\hat{Y}_{t+1}^0(M) - Y_{t+1}^0| = |e_{t+1}^*(M)|$ will follow the folded normal distribution with mean $\mu_f(M) = \sqrt{2/\pi} \sigma_M$ and variance $\sigma_f^2(M) = \sigma_M^2(1 - 2/\pi)$ (Leone et al. 1961). In other words

$$\begin{aligned} E(|e_t(A)|) &= \sqrt{2/\pi} \sigma_A \quad \text{and} \\ E(|e_t(B)|) &= \sqrt{2/\pi} \sigma_B \end{aligned} \quad (2)$$

Suppose that we score the performance of model M by

$|e_t(M)|$. Then, our selection will be based on the model with the minimum score. A natural choice of hypotheses to test could be:

$$\begin{aligned} H_0: E(|e_t(A)|) &= E(|e_t(B)|) \\ H_1: E(|e_t(A)|) &< E(|e_t(B)|) \end{aligned} \quad (3)$$

Because of (2) this is equivalent to

$$\begin{aligned} H_0: \sigma_A^2 &= \sigma_B^2 \\ H_1: \sigma_A^2 &< \sigma_B^2 \end{aligned} \quad (4)$$

Here we must note that

$$\text{Cov}(|e_t(A)| + |e_t(B)|, |e_t(A)| - |e_t(B)|) = (\sigma_A^2 - \sigma_B^2) \left(1 - \frac{2}{\pi}\right). \quad (5)$$

Set $R_{t+} = |e_t(A)| + |e_t(B)|$ and $R_{t-} = |e_t(A)| - |e_t(B)|$. Then, because of (5) the set of hypotheses (4), is equivalent to

$$\begin{aligned} H_0: \text{Cov}(R_{t+}, R_{t-}) &= 0 \\ H_1: \text{Cov}(R_{t+}, R_{t-}) &< 0 \end{aligned}$$

or to

$$\begin{aligned} H_0: \rho_{R_{t+}, R_{t-}} &= 0 \\ H_1: \rho_{R_{t+}, R_{t-}} &< 0. \end{aligned}$$

The obvious choice of a test statistic would be

$$R = \frac{\left[\frac{\sum R_{t+} R_{t-}}{n} - \bar{R}_{t+} \bar{R}_{t-} \right]}{\sqrt{\frac{\sum (R_{t+} - \bar{R}_{t+})^2}{n} \frac{\sum (R_{t-} - \bar{R}_{t-})^2}{n}}}$$

Then, under H_0 the asymptotic distribution of $\sqrt{n} R$ is normal with mean zero and variance $\text{Var}(R_{t+} R_{t-}) / \text{Var}(R_{t+}) \text{Var}(R_{t-})$ (Lehmann (1986)). So, values of $\sqrt{n} R$ in the left tail of the normal distribution will call for rejection of H_0 , thus indicating that model A performs better than model B.

References

- Lehmann, E. L. (1986). "Testing Statistical Hypotheses". J. Wiley and Sons.
- Leone, F.C. Nelson, L.S. and Nottigham, R.B. (1961). "The Folded Normal Distribution". *Technometrics*, 3(4), 543-550.
- Xekalaki E. and Katti, S.K. (1984). "A Technique for Evaluating Forecasting models". *Biom. J.*, 26, 173-184.