METEOROLOGICAL OFFICE
131415
8 FEB 1980
LIBRARY

MET O 11 TECHNICAL NOTE NO 135

Theil's Decompositions of Mean Square Error

1. Introduction

The shortcomings of the Mean Square Error (MSE) as an index of excellence or performance measure are as widely appreciated as the index is used. Theil's Decompositions do much to remove these shortcomings by providing an indication of the contributory components of the total MSE.

By

R. Dixon.

Met O 11 (Forecasting Research Branch)
Meteorological Office
London Road
Bracknell
Berks RG12 2QR, UK

Feb 1980

N.B. This paper has not been published. Permission to quote from it must be obtained from the Assistant Director of the above Meteorological Office Branch.

FH2A

2. Theil's Decompositions

I present the matter in the familiar context of a series of forecast values F_i which are verified against a series of corresponding actuals A_i, i = 1,----, N although it will be evident that the results apply to any two series in general.

Define

$$MSE = \frac{1}{N}\sum_1^N (F_i - A_i)^2 , \quad \bar{A} = \frac{1}{N}\sum_1^N A_i , \quad \bar{F} = \frac{1}{N}\sum_1^N F_i$$

$$s_A^2 = \frac{1}{N}\sum_1^N (A_i - \bar{A})^2 , \quad s_F^2 = \frac{1}{N}\sum_1^N (F_i - \bar{F})^2$$

$$r = \frac{\frac{1}{N}\sum_1^N (F_i - \bar{F})(A_i - \bar{A})}{s_F s_A}$$

respectively the Mean Square Error, the sample means, the sample variances (and thus the sample standard deviations s_F and s_A) and the sample correlation coefficient.

Henri Theil (1966) has pointed out that the MSE may be decomposed in various ways, two of which are

$$MSE = (\bar{F} - \bar{A})^2 + (s_F - s_A)^2 + 2(1 - r)s_F s_A \qquad (1)$$

and

$$MSE = (\bar{F} - \bar{A})^2 + (s_F - r s_A)^2 + (1 - r^2) s_A^2 \qquad (2)$$

3. The interpretation of Theil's components

The terms on the RHS of (1) and (2) have interesting and useful interpretations -

(a) $(\bar{F} - \bar{A})^2$ scarcely needs comment, being the component due to a simple mismatch of the mean levels of the two series. That is, even if the variance and covariance are perfect a contribution arises because the mean levels are different. In sinusoidal terms one may visualize a forecast curve which is exactly the same shape as the actual curve, same phase, same amplitude, but varying about a different mean.

(b) $(s_F - s_A)^2$ is a component due simply to unequal variation. The component arises because the variance of the forecast series is wrong. In sinusoidal terms one may picture a forecast curve which is varying about the same mean level as the actual curve, is exactly in phase, but has the wrong amplitude.

(c) $2(1 - r)s_F s_A$ is a component which arises if the forecast series does not vary in step with the actual series. In sinusoidal terms one may picture a forecast series which is varying about the same mean level as the actual series, has the same shape and amplitude but differs in phase from the actual series.

The interpretation of the remaining two terms (d) and (e) requires a change of viewpoint. Consider now that the F_i is a series of values which is available to act as the predictor variable in a linear regression relationship in which the series A_i is the predicted variable. The linear regression relationship will have the form

$$A = aF + b + \varepsilon \qquad (3)$$

where a and b are coefficients to be determined by least-squares and ε is a disturbance term. From elementary regression theory

$$a = r\frac{S_A}{S_F} \qquad \text{and} \qquad b = \bar{A} - r\frac{S_A}{S_F}\bar{F} \tag{4}$$

so that (3) becomes

$$A = r\frac{S_A}{S_F}F + \left(\bar{A} - r\frac{S_A}{S_F}\bar{F}\right) + \varepsilon \tag{5}$$

Now to isolate the contribution of the term $(S_F - rS_A)^2$ we assume that the series F_i has the correct mean level, so that $\bar{A} = \bar{F}$, and that there are no disturbance terms, so that $\varepsilon_i = 0$, i = 1,---, N. Eqn (5) then becomes

$$A = r\frac{S_A}{S_F}F + \left(1 - r\frac{S_A}{S_F}\right)\bar{F} \tag{6}$$

It now becomes apparent that if the mean level of F is right and there are no disturbances about a linear relationship then the forecasts will be exactly correct, $F_i = A_i$ for all i = 1,---, N, if a (i.e. $r\frac{S_A}{S_F}$) $= 1$.

(d) The interpretation of $(S_F - rS_A)^2$ now follows as it can be written as

$$(S_F - rS_A)^2 = S_F^2\left(1 - r\frac{S_A}{S_F}\right)^2 = S_F^2(1-a)^2 \tag{7}$$

It is thus the contribution to the total MSE which will arise from a departure of the linear regression coefficient a from 1 quite apart from any contribution from any disturbances or a wrong mean level.

(e) The contribution $(1-r^2)S_A^2$ is that which arises by virtue of the presence of disturbances that cannot be accounted for by a linear regression relationship. This follows immediately from regression theory since $(1-r^2)S_A^2$ is itself the expression for the residual variance, that part of the variance of A_i which cannot be explained by a simple linear regression on F

4. <u>Summary</u>

MSE	=	$(\bar{F}-\bar{A})^2$ +	$(S_F - S_A)^2$ +	$2(1-r)S_F S_A$
		MEAN LEVEL	VARIANCE (AMPLITUDE)	COVARIANCE (PHASE)

MSE	=	$(\bar{F}-\bar{A})^2$ +	$(S_F - rS_A)^2$ +	$(1-r^2)S_A^2$
		MEAN LEVEL	$a \neq 1$ IN LINEAR REGRESSION ON F	VARIANCE OF A NOT EXPLAINED BY LINEAR REGRESSION ON F

It is also useful to express the RHS terms as proportions of the MSE, thus displaying the relative importance of each contribution.

R Dixon

R Dixon
Met O 11

February 1980