The Comparison of Regression Variables

By E. J. Williams
North Carolina State College

[Received January, 1959]

SUMMARY

Hotelling's test of significance for difference in efficiency of predictors is reformulated in terms of regression analysis. A test proposed by Healy is shown to differ from Hotelling's test in general.

When a regression relationship for predicting the values of a variable is to be determined, and two or more regression variables have been measured, it will generally be appropriate to calculate the multiple regression to give the predictor. However, it sometimes happens that only one variable is to be used in the predictor, or that after some variables have been included, the choice of one out of the remaining variables is to be made. For the purpose of comparing the efficiency of different regression variables, Hotelling (1940) has developed a significance test, in terms of the correlations among the variables. Healy (1955) has discussed this test and shown that it can be derived from a regression analysis which makes the calculations simpler.

Healy's result, however, is in general different from Hotelling's. The present note derives Hotelling's test and expresses it as a test of regressions.

We suppose that we have a sample of n values of each of p regression variables $x_1, x_2, \ldots, x_p$, and an independent variable y. The test developed by Hotelling depends on the fact that the sum of squares for the simple regression of y on x_i is the square of a linear function of y; the null hypothesis is that each of these linear functions has the same expected value (thus, in particular, the null hypothesis will specify the sign as well as the magnitude of the correlation of y with each x_i).

We may write

$$z_i = Sy(x_i - \bar{x}_i)/\sqrt{S(x_i - \bar{x}_i)^2} = p_i/\sqrt{t_{ii}}$$

for the square root of the sum of squares for the regression of y on x_i. Then the variances and covariances of the z_i, in terms of the residual variance σ^2, are as follows:

$$V(z_i) = \sigma^2$$

$$\text{cov}\,(z_h, z_i) = r_{hi}\sigma^2,$$

where r_{hi} is the sample correlation of x_h and x_i.

Now let the elements of the inverse of the sample correlation matrix of the independent variables be denoted r^{hi}. Then the weighted mean of the z_i is given by

$$z = \sum_h \sum_i r^{hi} z_i / \sum_h \sum_i r^{hi}$$

and the sum of squares of the deviations of the z_i from their mean is

$$\begin{aligned}(p-1)\,s'^2 &= \sum_h \sum_i r^{hi} z_h z_i - (\sum_h \sum_i r^{hi} z_i)^2 / \sum_h \sum_i r^{hi} \\ &= \sum_h \sum_i r^{hi} z_h z_i - z^2 \sum_h \sum_i r^{hi}. \qquad (1)\end{aligned}$$

The sum of squares (1), with $p - 1$ degrees of freedom, provides a criterion for the reality of differences among the z_i, and consequently for the differences in efficiency among the x_i as predictors for y. It may be tested against the residual mean square, s^2, from the multiple regression of y on the x_i. Thus the ratio s'^2/s^2 is distributed as F with $p - 1$ and $n - p - 1$ degrees of freedom.

The above derivation is satisfactory for demonstrating algebraically the method of arriving at the significance test, but for calculation purposes it can be presented more simply. Rather than the inverse of the matrix $((r_{hi}))$ it is preferable to determine the inverse of the matrix $((t_{hi}))$ of sums of squares and products of the regression variables. This has the advantage of avoiding the calculation of the correlation coefficients; besides, the inverse of $((t_{hi}))$ is required in any case for determining the multiple regression equation and the sum of squares for multiple regression.

If the elements of the inverse of $((t_{hi}))$ are denoted by t^{hi}, then $r^{hi} = (t_{hh}t_{ii})^{\frac{1}{2}}\, t^{hi}$, and the terms in the sum of squares of the z_i, given above, may be expressed as follows:

$$\begin{aligned}\sum_h \sum_i r^{hi} z_h z_i &= \sum_h \sum_i t^{hi} p_h p_i \\ &= \sum_i b_i p_i \\ &= \text{sum of squares for multiple regression,}\end{aligned} \tag{2}$$

while

$$\begin{aligned}(\sum_h \sum_i r^{hi} z_i)^2 / \sum_h \sum_i r^{hi} &= [\sum_h t_{hh}^{\frac{1}{2}} \sum_i t^{hi} p_i]^2 / [\sum_h \sum_i (t_{hh} t_{ii})^{\frac{1}{2}}\, t^{hi}] \\ &= [\sum_h t_{hh}^{\frac{1}{2}} b_h]^2 / [\sum_h t_{hh}^{\frac{1}{2}} \sum_i t^{hi}\, t_{ii}^{\frac{1}{2}}],\end{aligned} \tag{3}$$

where the b_i are the partial regression coefficients.

The second term in (3) is seen to be the sum of squares for regression of y on the compound variate

$$x_0 = \sum_h t_{hh}^{\frac{1}{2}} \sum_i t^{hi} x_i.$$

Thus, framed in this way, the test is equivalent to a test of the adequacy of the compound variate x_0 as a regression function to replace the multiple regression function; x_0 would be equivalent to the multiple regression function if each of the variables x_i were equally highly correlated in the sample with y.

For computing purposes, the sum of squares for regression on x_0 is most conveniently represented in the form (3), in terms of the partial regression coefficients.

The test may be made from the following analysis of variance:

	Degrees of freedom	*Sum of squares*
Regression on x_0	1	$(\Sigma\, t_{ii}^{\frac{1}{2}} b_i)^2 / \Sigma\, t_{hh}^{\frac{1}{2}} \Sigma\, t^{hi} t_{ii}^{\frac{1}{2}}$
Differences in efficiency	$p - 1$	by difference $= (p - 1)\, s'^2$
Regression on $x_1, x_2, \ldots, x_p$	p	$\Sigma\, b_i p_i$
Residual	$n - p - 1$	$u - \Sigma\, b_i p_i = (n - p - 1)s^2$
Total	$n - 1$	u

We note that Healy eliminates, not the variate x_0, but a variate

$$x_0' = \sum_i x_i/\sqrt{t_{ii}}.$$

The sum of squares for regression of y on x_0' is

$$\frac{(\sum_i p_i t_{ii}^{-\frac{1}{2}})^2}{\sum_h \sum_i r_{hi}} = \frac{(\sum_i z_i)^2}{\sum_h \sum_i r_{hi}}. \tag{4}$$

When each of the x_i is equally correlated with y (each x_i an equally efficient predictor), $= z$, and the regression sum of squares, from (2), becomes

$$z^2 \sum_h \sum_i r^{hi},$$

while the sum of squares (4) for regression on x_0' becomes

$$\frac{p^2 z^2}{\sum_h \sum_i r_{hi}},$$

which is less than the multiple regression sum of squares. Hence, even in this case, the sum of squares for difference in efficiency, when determined from x_0', will differ from zero. This appears to be an unsatisfactory feature of the test proposed by Healy.

The limitations of Hotelling's test should be noted. In the first place, it is strictly a conditional test. It is valid for comparing the efficiency, as predictors, of the given sets of values of the independent variables, without reference to any population from which they might have been drawn. It cannot, however, validly be extended to drawing conclusions about future observed values of the independent variables.

In a few cases it may be reasonable to assume that the independent variables are fixed, and that the conditional test is all that is required. In some other cases it may be found that the test, while not exact, is a good approximation. In general, however, if a test of the efficiency of predictors for future use is required, some allowance will have to be made for the variation of the values from those observed. This question has not received much attention. An approximate test has been derived for comparing two predictors, when the variables are assumed drawn from a normal population. The test criterion, distributed approximately as F with 1 and $n - 3$ degrees of freedom, is

$$\frac{(z_1 - z_2)^2}{2s^2(1 - r_{12}) + \dfrac{(z_1 + z_2)^2(1 - r_{12})^3}{4(n-1)(1 + r_{12})}}, \tag{5}$$

where s^2 is the residual mean square from the joint regression. This may be compared with the criterion given by Hotelling's test, namely

$$\frac{(z_1 - z_2)^2}{2s^2(1 - r_{12})}. \tag{6}$$

The additional term in the denominator of (5) makes allowance for the variation in the x_i.