Misc Com

THE COMPARISON OF REGRESSION VARIABLES

Prepared under Contract No. DA-36-034-ORD-1517 (RD)

(Experimental Designs for Industrial Research)

by

E. J. Williams

Institute of Statistics

Mimeo. Series No. 195

April, 1958

THE COMPARISON OF REGRESSION VARIABLES

by

Evan J. Williams
North Carolina State College

I. Summary

Tests of significance for comparing the efficiency of different predictors in regression are considered. It is shown that a test of significance derived by Hotelling (1940) is valid only as a strictly conditional test, its conclusions being applicable only to the particular observed values of the predicting variables, and cannot be extended to the population of values from which the observed values of the predictors may have been drawn.

Alternative tests of significance applicable to tests of the wider hypothesis are considered, and one is examined in some detail.

II. Statement of the Problem

When a regression relationship is to be determined and there are measurements on several independent variables, it will generally be appropriate to calculate a relationship including all the variables, or at any rate all those that contribute materially to the relationship. However, sometimes a choice of one only of the regression variables is to be made, either for reasons of economy in subsequent applications of the relationship, or because the additional contribution of more than one variable is not expected to be important. Other things being equal, the variable chosen would be that for which the sum of squares for simple regression of the dependent variable would be largest. This variable might be expected to be the most efficient predictor. In such cases it is desirable to have a test whether the sums of squares for regression on different independent variables are significantly different.

III. Hotelling's Test

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 is expressed in terms of correlations, but will be presented here in terms of regression sums of squares. It 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, it should be noted that the hypothesis will specify the sign as well as the magnitude of the correlation of y with each x_i. Since the efficiency of a predictor does not depend on the sign of the correlation, this may be considered a shortcoming of the test, though it will be a characteristic also of alternative tests presented here.

We may write

$$u_i = S\, y(x_i - \bar{x}_i)/\sqrt{S(x_i-\bar{x}_i)^2}$$

$$= p_i/\sqrt{t_{ii}}$$

for the square root of the sum of squares for the regression of y on x_i. Hotelling considers the distribution of these quantities, conditional on fixed values of the x_i. In terms of the residual variance σ^2, the conditional variances and covariance of the u_i are

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

$$\text{Cov}\,(u_h, u_i) = r_{hi}\sigma^2$$

where r_{hi} is the sample correlation of x_h and x_i. Thus the conditional variance of any difference, for example $u_1 - u_2$, is

$$2\sigma^2(1 - r_{12}).$$

More generally, we may compare the set of p quantities u_i by determining the sum of squares for differences among them, taking into account the correlations among them. The validity of the conditional test is intuitive; but it will be helpful to the subsequent discussion to set it out formally.

We assume

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + e$$

where e is an independent random error, so that

$$E(u_i) = \sum_h \beta_h t_{hi} / \sqrt{t_{ii}},$$

where $$t_{hi} = S(x_h - \bar{x}_h)(x_i - \bar{x}_i).$$

The null hypothesis asserts that all these expectations have a common value, which we may denote by λ. The null hypothesis may also be expressed as a relationship satisfied by the partial regression coefficients, namely

$$\beta_i = \lambda \sum_h t^{hi} \sqrt{t_{hh}}$$

$$= \lambda \sum_h r^{hi} / \sqrt{t_{ii}}$$

where the r^{hi} are the elements of the inverse of the sample correlation matrix of the x_i.

Regression coefficients satisfying these conditions will be obtained if the regression of y on the compound variate

$$X = \sum_h \sum_i \sqrt{t_{hh}} . t^{hi} x_i$$

is determined.

Then the test of significance for the difference of efficiency of the different predictors x_i may be set out in the following analysis of variance:

	Degrees of freedom
Regression on X	1
Regression on remaining variables	p - 1
Residual	n - p - 1
Total	n - 1

The mean square for regression on the remaining variables with p - 1 degrees of freedom is tested against the residual mean square. It may be shown that this test is equivalent to Hotelling's test. The variable X is so constructed that it will agree with the multiple regression function, and its sum of squares with the multiple regression sum of squares, provided each of the variables x_i is equally highly correlated in the sample with y.

The advantages of the regression approach to this problem are believed to be, firstly, that it shows up more clearly the nature of the test and the precise form of the null hypothesis which is being tested, and secondly, that the computations are simpler, being an extension of the usual multiple regression analysis.

IV. Tests of a Wider Hypothesis

We shall often be interested in testing whether variables to be used as predictors differ significantly in their efficiency when their values are not restricted to those observed and used in making the test. In making this wider test we have to assume that the values observed are in fact representative of the population from which future samples will be drawn (an assumption which is not always satisfied), and that the distributional form in the population is known. Needless to say, the wider test will be less sensitive than the conditional test;

the effect of assuming the x_i to be distributed is, roughly speaking, to increase the residual variance appropriate for making the test. No exact test of this hypothesis is known.

We shall assume that y and the x_i are drawn at random from a normal multivariate population, for which the above given regression model holds. We shall also assume without loss of generality that the variances of the x_i are unity, and denote the correlations by ρ_{hi}. Then the wider null hypothesis is expressed in terms of the variances and covariances of the x_i, rather than in terms of their sums of squares and products. It is in fact

$$\sum_h \beta_h \rho_{hi} = \lambda \qquad (i = 1, 2 \ldots p)$$

or

$$\beta_h = \lambda \sum_i \rho^{hi}$$

We shall here consider only the case p = 2, which shows sufficiently well the difficulties of the problem. For this case, the null hypothesis reduces to

$$\beta_1 = \beta_2 \ (= \beta, \text{ say}).$$

We first show that the conditional test is not valid for this hypothesis, even in the limit when the sample size is large. The expected values of u_1 and u_2, conditional on fixed values of the x_i, are

$$E_c(u_1) = \beta(\sqrt{t_{11}} + r\sqrt{t_{22}})$$

and

$$E_c(u_2) = \beta(r\sqrt{t_{11}} + \sqrt{t_{22}})$$

where we now, without risk of confusion, replace r_{12} by r. The expected value of the difference is, then

$$E_c(u_1-u_2) = \beta(\sqrt{t_{11}} - \sqrt{t_{22}})(1-r)$$

Although $t_{11}^{1/2}$ and $t_{22}^{1/2}$ are $O(n^{1/2})$, their difference is $O(1)$, so the conditional expected value of u_1-u_2 is $O(1)$. Since the variance of u_1-u_2 is $O(1)$, this result shows that the conditional test has in general a bias which does not tend to zero as the sample size tends to infinity.

V. Marginal Distribution of the Difference

Although the conditional expected value of u_1-u_2 does not vanish, it is easily seen, by symmetry, that its marginal expected value does. It might therefore be made the basis of a test based on its marginal distribution, for values of x allowed to vary. For such a test we require the marginal variance of the difference, or something equivalent.

Now for any statistic w, it can be shown that

$$V(w) = E(V_c(w)) + V(E_c(w)).$$

For present purposes it will be convenient to take

$$w = \frac{u_1-u_2}{1-r}$$

Then $$E_c(w) = \beta(\sqrt{t_{11}} - \sqrt{t_{22}})$$

and $$V_c(w) = \frac{2\sigma^2}{1-r}$$

so that $$V(w) = 2\sigma^2 E(\frac{1}{1-r}) + \beta^2 E(t_{11} - 2\sqrt{(t_{11}t_{22})} + t_{22})$$

The second term is clearly O(1); and, being both independent of scale and symmetric in the variances of x_1 and x_2 (the population variances in the expression having been suppressed), it may be shown to be an even function of ρ, the correlation of x_1 and x_2. We may determine its value in the following way. Clearly

$$E(t_{11} + t_{22}) = 2(n-1)$$

so that we need discuss only the evaluation of $E(\sqrt{(t_{11}t_{22})})$. The joint probability element of t_{11}, t_{22} and r is (Fisher 1915)

$$\text{const. } (1-\rho^2)^{-\frac{1}{2}(n-1)} \quad \exp\left\{\frac{q(\cosh z - \rho r)}{1-\rho^2}\right\} q^{n-2}\, dq\,dz\,dr\, (1-r^2)^{\frac{1}{2}(n-4)}$$

where

$$q^2 = t_{11}t_{22} \quad , \quad 2z = \log(t_{11}/t_{22})$$

and the joint probability element of z and r is

$$k\, (1-\rho^2)^{\frac{1}{2}(n-1)} \quad \frac{dz}{(\cosh z - \rho r)^{n-1}} \quad (1-r^2)^{\frac{1}{2}(n-4)}\, dr$$

Hence $E(\sqrt{(t_{11}t_{22})}) = E(q)$

$$= k(n-1)\, (1-\rho^2)^{\frac{1}{2}(n+1)} \int_{-1}^{+1} \int_{-\infty}^{+\infty} \frac{dz}{(\cosh z - \rho r)^n} (1-r^2)^{\frac{1}{2}(n-4)}\, dr$$

and on integrating by parts with respect to r we get

$$k(n-4)\, \frac{(1-\rho^2)^{\frac{1}{2}(n+1)}}{\rho} \int_{-1}^{+1} \int_{-\infty}^{+\infty} \frac{dz}{(\cosh z - \rho r)^{n-1}} (1-r^2)^{\frac{1}{2}(n-6)}\, r\, dr$$

$$= (n-4)\, \frac{1-\rho^2}{\rho}\, E\left(\frac{r}{1-r^2}\right)$$

Now expanding $\frac{r}{1-r^2}$ in powers of $\theta = r-\rho$:

$$\frac{r}{1-r^2} = \frac{\rho}{1-\rho^2} + \frac{\theta(1+\rho^2)}{(1-\rho^2)^2} + \frac{\theta^2(3\rho+\rho^3)}{(1-\rho^2)^3} + \ldots.$$

and using the results given by Hotelling (1953, page 212) we have

$$E\left(\frac{r}{1-r^2}\right) = \frac{\rho}{1-\rho^2}\left(1 + \frac{5+\rho^2}{2(n-1)} + \frac{61 + 10\rho^2 + \rho^4}{8(n-1)^2} + \ldots..\right)$$

so that

$$E(q) = (n-4)\left(1 + \frac{5+\rho^2}{2(n-1)} + \frac{61 + 10\rho^2 + \rho^4}{8(n-1)^2} + \ldots..\right)$$

Hence

$$E(t_{11} - 2\sqrt{(t_{11}t_{22})} + t_{22}) = 2(n-1 - E(q))$$

$$= 1-\rho^2 - \frac{(1-\rho^2)^2}{4(n-1)} + O(n^{-2})$$

We therefore have

$$V(w) = 2\sigma^2 E\left(\frac{1}{1-r}\right) + \beta^2(1-\rho^2) + O(n^{-1}),$$

which exceeds the expected value of the conditional variance of w by a term $O(1)$.

It does not appear possible to determine an unbiased estimate of this variance. An estimate which is independent of w, and whose bias is $O(n^{-1})$, is

$$\frac{2s^2}{1-r} + \frac{(u_1+u_2)^2(1-r)}{4(n-1)(1+r)} .$$

VI. Discussion of the Alternative Hypotheses

We have shown above that Hotelling's test is appropriate for testing the limited hypothesis of equal efficiency when the predictors are held fixed, but that, as a test of the wider hypothesis $\rho_1 = \rho_2$, it is biased, even for large samples. Viewed as a conditional test, it would be in error since the quantities being compared would not in general have the same expected value; while viewed as an unconditional test, it would fail because the variance used to test the difference is the conditional variance and therefore too small.

The fact that Hotelling has phrased the test in terms of correlations may have been misleading to some users, for the correlation approach is usually and appropriately only associated with random sampling from a multivariate population, rather than with a relationship to fixed values of the independent variables. Indeed, it is probably this phrasing in terms of correlations that has led to the impression sometimes held that the test is inexact.

The above discussion shows that the test is valid for a specific null hypothesis, so that the only point at issue is what null hypothesis is appropriate. Indeed, it may be questioned whether the null hypothesis, for which Hotelling's test is valid, is ever appropriate. For the implication of the test is that, even when y is dependent on p variables $x_1, x_2, \ldots, x_p$ which are fixed and not random, it could be appropriate to use for prediction purposes the regression of y on any one of the x_i.

VII. Alternative Tests

We should like to find some suitable alternative test for comparing the efficiency of different predictors. Since if a <u>similar</u> test criterion (i.e. one whose distribution is independent of any nuisance parameters) exists, it is usually

fairly readily found, it seems that no such test exists in this problem. It is therefore necessary to compare the advantages of various approximate alternatives. Some of these will now be considered.

1. Comparison of regression sums of squares.

The most appropriate criterion for practical purposes is the difference of the regression sums of squares, for the magnitude of this difference is what is of interest to the experimenter. It appears from the tone of Hotelling's paper that this criterion was considered by him and abandoned in favour of the difference of correlation coefficients. It is therefore a fair presumption that the difference of regression sums of squares is troublesome to work with. Being an indefinite quadratic form of rank 2 and signature zero, it would not be amenable to any familiar form of analysis.

2. Comparison of correlations.

The second possibility would be to study the difference of correlations and to endeavour to find an approximation to its marginal distribution. This has already been attempted in Section IV. We there worked, not with differences of correlations as such, but with

$$w = \frac{u_1 - u_2}{1-r}$$

for mathematical convenience. All such multiples of the difference $u_1 - u_2$ would of course be equivalent for the conditional distribution, but some might have simpler marginal distributions than others. It has been found, for instance, that the marginal variance of w is easier to determine than that of $u_1 - u_2$.

In passing, it may be mentioned that the test of the simple correlations is equivalent to that on the partial correlations. For suppose we denote the simple

correlation coefficients by r_1 and r_2, and the corresponding partial coefficients by r_1' and r_2'. Then, since

$$r_1' = \frac{r_1 - r_2 r_{12}}{\sqrt{(1-r_{12}^2)}} \quad , \quad \text{etc.}$$

we have

$$r_1' - r_2' = (r_1 - r_2) \sqrt{\left(\frac{1+r_{12}}{1-r_{12}}\right)} \quad ,$$

so that the ratio of $r_1' - r_2'$ to $r_1 - r_2$ depends only on r_{12}. This result does not carry over to transformations of the coefficients.

3. Comparison of Transformed Correlations.

Since Fisher has shown that a hyperbolic transformation of the correlation coefficient makes the form of its distribution practically independent of the population correlation, it would seem that some such transformation might be effective for the present problem. However, since the correlation of r_1 and r_2 will, on the null hypothesis, usually be large, the hyperbolic transformation is not appropriate, as it would be if we were comparing independent correlation coefficients. Hotelling shows that a large-sample approximation to the variance of the difference $r_1 - r_2$ is

$$(1-\rho)(2 - 3\rho_1^2 + \rho\rho_1^2)/n$$

where ρ_1 is the correlation of y with x_1 or x_2, and ρ is, as before, the correlation of x_1 and x_2.

From this form for the variance, we can deduce the appropriate transformation for the correlations; the transformed variables would be

$$v_i' = \arcsin r_i \sqrt{\left(\frac{3-\rho}{2}\right)}$$

Since the argument of the inverse sine may sometimes exceed unity owing to sampling fluctuations, the simpler transformation

$$v_i = \arcsin r_i$$

might be more practical. Some of the properties of this transformation have already been studied by Harley (1956, 1957) who shows in particular that

$$E(\arcsin r_i) = \arcsin \rho_1 \quad .$$

We shall be concerned mainly to find a transformation such that the variance of the difference of the transformed values is practically independent of ρ_1, expecting that the variance will be a function of ρ which can be estimated from some function of r.

If this transformation is applied, we shall find that

$$V(v_1-v_2) = (1-\rho)\left(2 - \frac{\rho_1^2(1-\rho)}{1-\rho_1^2}\right)/n$$

approximately. This transformation is not effective in rendering the variance of the difference relatively stable; in this problem it seems difficult to find any transformation which would do so.

4. The Likelihood Ratio Test

Any test of the efficiency of two predictors x_1 and x_2 must be based on the three sample correlation coefficients r_1, r_2 and r. The joint distribution of these coefficients depends on the corresponding population correlation coefficients ρ_1, ρ_2 and ρ. Since the null hypothesis specifies merely that $\rho_1 = \rho_2$ without specifying the values of any of these parameters, the hypothesis is composite. A general criterion for testing significance in such cases is based on the ratio of likelihood maximized subject to the restriction of the null hypothesis to that

maximized without restriction on the parameters. In this case, minus twice the logarithm of such a ratio would be distributed asymptotically as χ^2 with 1 degree of freedom.

As the restricted maximum likelihood is difficult to evaluate, this approach has not been followed here; we hope to work on this test later.

VIII. Conclusions

From the above discussion it appears that the simplest and most effective test of the difference of efficiency of predictors is that given at the end of Section IV: this is, in effect, a test of the difference of standardized regression coefficients against its marginal standard error.

Further investigation is needed, of

(a) alternative test criteria, and

(b) generalizations of the test to the comparison of more than two predictors.

IX. Acknowledgement

I am indebted to William H. Kruskal for bringing this problem to my notice, and for helpful discussions.

X. References.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 10, 507-521.

Harley, B. I. (1956). Some properties of an angular transformation of the correlation coefficient. Biometrika 43, 219-224.

Harley, B. I. (1957). Further properties of an angular transformation of the correlation coefficient. Biometrika 44, 273-275.

Hotelling, H. (1940). The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. Ann. Math. Stat. 11, 271-283.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. Jour. Roy. Stat. Soc. B 15, 193-225.