# NOTES

## An Average Correlation Coefficient

ALAN J. FALLER

*Institute for Physical Science and Technology, University of Maryland, College Park 20742*

3 February 1980 and 22 September 1980

ABSTRACT

It is proposed that when a dependent variate $y$ is classified into $N$ sets of values with different predictive statistical relations for each classification, the associated correlation coefficients $r_n$ ($n = 1$–$N$) can be usefully combined into a single correlation coefficient by a weighted average of the $r_n^2$. The weighting factor is the variance of $y$ for each $n$. In assessing the combined predictability of $y$ by the set of $N$ relations an additional weighting factor is $J_n$, the number of predictions to be made for each $n$.

### 1. Introduction

Occasions sometimes arise where for the presentation of statistics in a concise form it would be desirable to give an average correlation coefficient rather than a large number of individual values. Let us suppose, for example, that the price of tea in China is linearly related to the temperature in Peking and the wind speed in Canton. We imagine that we can predict the price of tea by first forecasting the weather in Peking and Canton and by then using a regression equation. As often happens, however, it is found to be desirable to classify the data and to use a different regression equation for each month of the year, and in certain months predictors other than those mentioned above might be found preferable. Thus, 12 distinct regression equations would be formulated, and each equation would have an associated multiple-correlation coefficient $r_n$, $n = 1 - N, N = 12$. The success of each regression in reducing the variance in the observed price of tea would be measured by $r_n^2$, the square of the multiple-correlation coefficient, also known as the fractional reduction of variance. It would be desirable to present an overall or average multiple-correlation coefficient $R$, or an overall reduction of variance $R^2$, to measure the success of all of the regression equations taken together. A definition of a suitable $R^2$ is given below and it is shown that this definition is equivalent to a weighted average of the $r_n^2$.

The above problem, $N = 12$, may seem trivial, but an occasion has arisen in research with the statistical correction of numerical predictions (Schemm *et al.*, 1981) where 128 separate regression equations were calculated, one for each grid point on a 16 × 8 grid. Moreover, it was found necessary to compare several different statistical procedures, each having 128 regression equations. Thus an overall measure of the success of each procedure was desirable for comparison of the several procedures. With the continued rapid growth of computer capability and the proliferating use of regression equations in many fields, the concept of an average correlation coefficient may find increasing application.

### 2. Some elementary considerations

The fundamental definition of the correlation coefficient may be written as (Hoel, 1947)

$$r^2 = 1 - \frac{s_e^2}{s_y^2}, \tag{1}$$

where $s_y^2 = \overline{(y_i - \bar{y}_i)^2}$ is the sample variance of the dependent variate $y$, and the overbar indicates the sample mean, $i = 1 - I$. The numerator on the right of (1) is $s_e^2 = \overline{(y_i - \hat{y}_i)^2}$, where $\hat{y}_i$ is an estimate of $y_i$ and is called the residual variance. The estimates $\hat{y}_i$ are usually associated with a regression equation, and in the case of linear regression with a single independent variate $x$ the regression equation would be

$$\hat{y}_i = a + bx_i, \tag{2}$$

where $a$ and $b$ are the regression coefficients. Then $r$ is given by the familiar relation

$$r = \frac{\overline{(x_i - \bar{x}_i)(y_i - \bar{y}_i)}}{s_x s_y}. \tag{3}$$

It should be recognized, however, that the definition of $r$ in (1) is not restricted to the use of linear regression equations and, in fact, any suitably codified method of producing the estimates $\hat{y}_i$ should be acceptable, the resulting value of $r^2$ then corresponding to that method.

When many such regression equations (or prediction methods) are used for classified data, the reduction of variance associated with the $n$th regression

0021-8952/81/020203-03$04.75

equation is

$$r_n^2 = 1 - \frac{s_{e_n}^2}{s_{y_n}^2}, \tag{4}$$

where

$$s_{y_n}^2 = I_n^{-1} \sum_{i=1}^{I_n} (y_{in} - \bar{y}_{in}^n)^2,$$

$$s_{e_n}^2 = I_n^{-1} \sum_{i=1}^{I_n} (y_{in} - \hat{y}_{in}^n)^2, \tag{5}$$

and $I_n$ is the number of samples in the $n$th case. Here it is to be understood that the symbols $\hat{y}_{in}^n$ and $\bar{y}_{in}^n$ refer, respectively, to estimates made using the $n$th regression equation and to means for the corresponding sample values of $y$. These estimates and means are to be distinguished later from estimates $\hat{y}_{in}$ using a single regression equation for all data and the grand mean $\bar{y}_{in}$ for all data. For classified data with $N$ linear regression equations and a single independent variate in each equation, the $n$th equation would be

$$\hat{y}_{in}^n = a_n + b_n x_{in}, \tag{6}$$

where it should be noted that the $x_{in}$ can represent a different predictor for each $n$.

### 3. An overall correlation coefficient

We now define an overall multiple-correlation coefficient by

$$R^2 \equiv 1 - \frac{\sum_{n=1}^{N} s_{e_n}^2}{\sum_{n=1}^{N} s_{y_n}^2}. \tag{7}$$

To motivate the form of (7), it is useful to write (4) as

$$s_{e_n}^2 = (1 - r_n^2) s_{y_n}^2 \tag{8}$$

and to sum over $n$. The result is

$$\sum_{n=1}^{N} s_{e_n}^2 = \sum_{n=1}^{N} (1 - r_n^2) s_{y_n}^2. \tag{9}$$

Then by analogy with (8) the form of (9) suggests that it might be rewritten as

$$\sum_{n=1}^{N} s_{e_n}^2 = (1 - R^2) \sum_{n=1}^{N} s_{y_n}^2 \tag{10}$$

which reduces directly to (7). It also may be seen from (9) and (10) that

$$R^2 = \frac{\sum_{n=1}^{N} r_n^2 s_{y_n}^2}{\sum_{n=1}^{N} s_{y_n}^2}, \tag{11}$$

illustrating the fact that $R^2$ as defined in (7) is also an average of the $r_n^2$ weighted by the $s_{y_n}^2$.

### 4. Examples

We first consider a case with $N = 2$ and with the contrived values

(Set 1) $n = 1$, $s_{y_1}^2 = 1.00$,

$s_{e_1}^2 = 0.10$ $(r_1^2 = 0.90)$,

(Set 2) $n = 2$, $s_{y_2}^2 = 0.10$,

$s_{e_2}^2 = 0.09$ $(r_2^2 = 0.10)$.

Application of (7) gives $R^2 = 1 - (0.19/1.10) = 0.83$. This high value of $R^2$ is appropriate because the set with the larger variance ($n = 1$) had a high correlation.

As a second example, consider the values

(Set 1) $n = 1$, $r_1^2 = 0.10$,

$s_{y_1}^2 = 1.00$ $(s_{e_1}^2 = 0.90)$,

(Set 2) $n = 2$, $r_2^2 = 0.90$,

$s_{y_2}^2 = 0.10$ $(s_{e_2}^2 = 0.01)$.

Application of (11) gives $R^2 = (0.19/1.10) = 0.17$. In this case $R^2$ is approximately small and close to the value $r_1^2 = 0.10$.

In each of these examples, a direct average would be $\overline{r_n^2} = 0.50$ and there would be no distinction between the first example where data with high variance were well explained and the second example where the high variance data were poorly explained.

### 5. Discussion

The definition (7) is somewhat arbitrary and requires justification by comparison with other possibilities. Two alternatives that come to mind serve as a focus for several important considerations.

A possible definition, one that was in fact advocated by certain colleagues until its deficiencies were clarified, is

$$R_1^2 = 1 - \frac{\sum_{n=1}^{N} \sum_{i=1}^{I_n} (y_{in} - \hat{y}_{in}^n)^2}{\sum_{n=1}^{N} \sum_{i-1}^{I_n} (y_{in} - \bar{y}_{in})^2}, \tag{12}$$

where the numerator is the sum of squares of all errors of estimate using the $N$ regression equations and the denominator is based on departures of the $y_{in}$ from the grand average $\bar{y}_{in}$ rather than from the several $\bar{y}_{in}^n$. Eq. (12) is a poor choice for two reasons. First, if we temporarily restrict considerations to cases with $I_n$ the same for all $n$, Eq. (12) can be written as

$$R_1^2 = 1 - \frac{\sum_{n=1}^{N} s_{e_n}^2}{\sum_{n=1}^{N} s_{y_n}^2 + K}, \tag{13}$$

where

$$K = \sum_{n=1}^{N} (\bar{y}_{in}^{n})^2 - N\bar{y}_{in}{}^2$$

is $N$ times the variance of $\bar{y}_{in}^{n}$. To obtain (13) both numerator and denominator of (12) were divided by the constant $I_n$ and the denominator was expanded into that of (7) plus additional terms. Except in the trivial case where all $\bar{y}_{in}^{n}$ are equal ($K = 0$), $K$ is always positive and $R_1{}^2 > R^2$. To see that (7) is the more reasonable formulation, consider the situation where $s_{e_n}^2 = s_{y_n}^2$, and thus $r_n = 0$ for each $n$. Then $R^2 = 0$ but

$$R_1{}^2 = \frac{K}{\sum_{n=1}^{N} s_{y_n}^2 + K} . \tag{14}$$

Since there is no reason why $K$ cannot equal or exceed $\sum_{n=1}^{N} s_{y_n}^2$ it is clear that the use of $\bar{y}_{in}$ in the denominator of (12) is inappropriate as well as inconsistent.

The second deficiency in (12) arises in cases with different $I_n$. If the denominator in (12) is corrected to

$$\sum_{n=1}^{N} \sum_{I=1}^{I_n} (y_{in} - \bar{y}_{in}^{n})^2,$$

the modified (12) may be written as

$$R_2{}^2 = 1 - \frac{\sum_{n=1}^{N} I_n s_{e_n}^2}{\sum_{n=1}^{N} I_n s_{y_n}^2} \tag{15}$$

or in the form of (11)

$$R_2{}^2 = \frac{\sum_{n=1}^{N} r_n{}^2 I_n s_{y_n}^2}{\sum_{n=1}^{N} I_n s_{y_n}^2} . \tag{16}$$

Eq. (16) differs from (11) by the additional weighting factor $I_n$, which is unacceptable. Imagine, for instance, that in the examples of Section 4 $r_1$ was based on $I_1 = 10^4$ random samples and $r_2$ on $I_2 = 10^5$ samples, and that both were accurate measures of their population correlations, $\rho_1$ and $\rho_2$. Clearly, weighting $r_2{}^2$ with a factor 10 times that for $r_1{}^2$ would be inappropriate.

Suppose, on the other hand, that in the *use* of our two regression equations 100 predictions were to be made with Set 1 ($J_1 = 100$) and only 50 predictions ($J_2 = 50$) were to be made using Set 2. Then a measure of the combined predictability would logically require a weighting by $J_n$, the number of predictions to be made using the $n$th equation. It is therefore appropriate to define a measure of the overall predictability associated with (7) and (11) by the relations

$$R^{*2} = 1 - \frac{\sum_{n=1}^{N} J_n s_{e_n}^2}{\sum_{n=1}^{N} J_n s_{y_n}^2} , \tag{17}$$

$$R^{*2} = \frac{\sum_{n=1}^{N} r_n{}^2 J_n s_{y_n}^2}{\sum_{n=1}^{N} J_n s_{y_n}^2} . \tag{18}$$

If equal numbers of predictions are to be made with each equation, (17) and (18), of course, reduce to (7) and (11).

A distinctly different average correlation coefficient is that associated with Fisher's $z$ transformation (Hoel, 1947)

$$z_n = \tfrac{1}{2} \ln\left(\frac{1 + r_n}{1 - r_n}\right) . \tag{19}$$

In suitably prescribed circumstances $z_n$ is approximately normally distributed with mean

$$m_z = \tfrac{1}{2} \ln\left(\frac{1 + \rho}{1 - \rho}\right) ,$$

where $\rho$ is the population correlation coefficient, and with standard deviation

$$\sigma_z = \frac{1}{\sqrt{I_n - 3}} .$$

Thus, if one sampled the same population $N$ times, an appropriate estimate of the true mean $m_z$ would be

$$\bar{z}_n = N^{-1} \sum_{i=1}^{N} z_n.$$

Then, by the inversion of (19) one would determine an average

$$\bar{r}_n = \frac{e^{2\bar{z}}n - 1}{e^{2\bar{z}}n + 1} .$$

Such an average, however, has a distinctly different function and meaning than that of (7) since all samples are drawn from the same population and have the same sample size. Moreover, the algebraic sign of $r_n$ is accounted for in the average.

By contrast, the recommended definitions (7) and (17) or their equivalents (11) and (18) are intended to measure the combined effectiveness of several correlations (not necessarily linear) with the same dependent variate but with several different sets of independent variates. Moreover, by averaging the squares $r_n{}^2$, negative and positive correlations have equal measure and do not tend to cancel.

*Acknowledgments.* This research has been supported in part by the National Science Foundation under Grants ATM 76-82061 and ATM 7924544.

## REFERENCES

Hoel, P. G., 1974: *Introduction to Mathematical Statistics.* Wiley, 258 pp.

Schemm, C. E., D. A. Unger, and A. J. Faller, 1980: Statistical corrections to numerical predictions, III. *Mon. Wea. Rev.*, **108**, 96–109.