ED 331 706 SE 051 972

AUTHOR	O'Brien, Francis J., Jr.
TITLE	A Derivation of the Limits of the Sample Multivariate Correlation Coefficient.
PUB DATE	Mar 91
NOTE	18p.
PUB TYPE	Guides - Classroom Use - Instructional Materials (For Learner) (051)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Algebra; *College Mathematics; *Correlation; Higher Education; Learning Activities; Mathematical Applications; Mathematics Education; *Multivariate Analysis; *Problem Solving; *Proof (Mathematics); *Statistics

ABSTRACT

This paper is the sixth in a series designed to supplement the statistics training of students. The intended audience is social science undergraduate and graduate students studying applied statistics. The purpose of the applied statistics monographs is to provide selected proofs and derivations of important relationships or formulas that students do not find available and/or comprehensible in journals, textbooks and similar sources. Derived is the theoretical limits of the sample multivariate (or multiple) correlation of one criterion (dependent variable) and any (finite) number of predictors (independent variables). The proof given in this paper involves deriving the individual terms of R. The lower limit and upper limit of R are derived separately. (KR)

A DERIVATION OF THE LIMITS OF THE SAMPLE MULTIVARIATE CORRELATION COEFFICIENT

Francis J. O'Brien, Jr., Ph.D.

36 Linden Street
Middletown, RI 02840
March, 1991

Table of Contents

List of Tables

A Derivation of the Limits of the Sample Multivariate Correlation Coefficient

Francis J. O'Brien, Jr., Ph.D.

Introduction

This paper is the sixth in a series of ERIC publications designed to supplement the statistics training of students. For related documents see O'Brien (1982a; 1982b; 1982c; 1984; 1987). The intended audience for these papers is social science undergraduate and graduate students studying applied statistics.

The purpose of these applied statistics monographs is to provide selected proofs and derivations of important relationships or formulas that students do not find available and/or comprehensible in journals, textbooks and similar sources. For example, based on the author's personal experience as a former applied statistics instructor at the graduate level, few students would profit from a reading of Kendall and Stuart (1967) to understand the proof provided in the present paper. The unique feature of the papers in this series is detailed step-by-step proofs or derivations written in a consistent notation system. Calculus is neither used nor assumed. Each proof or derivation is presented algebraically in detail.

The present paper assumes familiarity with the authors' 1982c paper (or equivalent knowledge). That paper formulated a detailed derivation of the sample multiple correlation formula for one dependent variable and p predictors for the linear model based on standardized (z) variables.

Introduction to Proof

In this paper we derive the theoretical limits of the sample multivariate (or multiple) correlation of one criterion (dependent variable) and any (finite) number of predictors (independent variables). To facilitate the development of the proof, we will work with standardized (z) variables. Although the proof could be presented in the unstandardized ("raw score") form, normalized variables reduces some of the algebraic details.

Many students have learned that the multivariate correlation between one dependent variable and a finite number of independent variables can be expressed as a weighted sum of regression weights and Pearson (zero-order) product-moment correlations between dependent/independent variables. This relationship holds only for standardized variables. This correlation for *p* independent variables can be written (see O'Brien, 1982c):

$$R_{z_Y \cdot z_1, z_2, \ldots, z_j, \ldots, z_p} = \sqrt{B_1 r_{y1} + B_2 r_{y2} + \ldots + B_j r_{yj} + \ldots + B_p r_{yp}} \; .$$

Writing the right-hand side in summation notation,

$$R_{z_Y \cdot z_1, z_2, \ldots, z_j, \ldots, z_p} = \sqrt{\sum_{j=1}^{p} B_j r_{yj}} \; .$$

where

$R_{z_Y \cdot z_1, z_2, \ldots, z_j, \ldots, z_p}$ = multiple correlation of *p* standardized variables,

z_Y = the standardized dependent variable

$z_1, z_2, \ldots, z_j, \ldots, z_p$ = the standardized independent variables

$B_1, B_2, \ldots, B_j, \ldots, B_p$ = beta (regression weights) attached to each standardized independent variable*

$r_{y1}, r_{y2}, \ldots, r_{yj}, \ldots, r_{yp}$ = product moment (zero-order) dependent/independent variable correlations.

Many students know that the numerical limits on the above multiple R are zero and 1 (i.e., $0 \leq R \leq 1$). The purpose of this paper is to prove that statement.

Proof that $0 \leq R \leq 1$

In this section we present a detailed proof that the limits of the multiple R are 0/1. First, a review is given of the notation and necessary definitions as well as the relevant results that were derived in O'Brien (1982c).

We can state the formal linear regression prediction equation for *p* standardized predictors as follows:**

$$Z_{\hat{Y}_i} = B_1 Z_1 + B_2 Z_2 + \ldots + B_j Z_j + \ldots + B_p Z_p$$

This equation represents the predicted standardized criterion measure or score ($Z_{\hat{Y}_i}$) for the *i*th subject in the sample on the *p* standardized variables Z_1 through Z_p.

* Technically, the beta weights (B_j) are called "standardized partial regression cofficients". The formal notation in some standard textbooks is more elaborate than ours (e.g., Hays, 1973 or Kendall and Stuart, 1967). As in previous papers, we have minimized the reading of the symbolism to clarify the concepts in the development of the proof.

** The coefficient "A" is not included for the reason given in O'Brien (1982c) ; i.e., it "drops out" in the least squares derivations and so may be ignored.

The multiple correlation (or just R for short) for this regression model of *p* standardized predictors may be defined conceptually as:

$$R \quad = \mathit{Corr}(Z_Y, Z_{\hat{Y}}) = \frac{\mathit{Cov}(Z_Y, Z_{\hat{Y}})}{\sqrt{\mathit{Var}(Z_Y)\mathit{Var}(Z_{\hat{Y}})}},$$

where *Corr* is the correlation operator, *Cov* is the covariance operator, and *Var* is the variance operator. Note that Z_Y is the random variable that represents the "observed" or known information while $Z_{\hat{Y}}$ represents the "predicted information".

The proof that is given in this paper involves deriving the individual terms of R. Two tables are provided for reference in the development of the proof. Table 1 summarizes familiar formulas for standardized variables. Table 2 is a summary of the results derived in O'Brien (1982c) for the multiple R of *p* standardized variables. The information in each table provides the essential building blocks of the proof.

Table 1

Formulas and Relationships for Sample Standardized Variables

Name of Quantity	Formula	Note
Sum	$\sum_{i=1}^{n} Z_i = 0$	n is sample size. The summation is undeıstood to be across the sample for a given predictor *j*.
Sum of Squares	$\sum_{i=1}^{n} Z_i^2 = n-1$	Above note applies.
Mean	$\frac{\sum_{i=1}^{n} Z_i}{n} = \bar{Z}_j = 0$	Mean of *j*th predictor for total sample. The summation is understood to be across the sample for a given predictor *j*.
Variance	$\frac{\sum_{i=1}^{n} Z_i^2}{n-1} = Var(Z_j) = 1$	Variance of *j*th predictor for total sample. The summation is understood to be across the sample for a given predictoı *j*.

(Table 1 cont.)

Correlation	$\frac{\sum_{i=1}^{n} Z_{X_i} Z_{Y_i}}{n-1} = r_{Z_X Z_Y} = r_{xy}$	General zero-order correlation formula for any two standardized variables, Z_X and Z_Y.

Note: Proof of these formulas/relationships may be found in O'Brien (1982b, Ap: endix).

Table 2

Formulas and Relationships for the Sample Multiple R

$$Cov(Z_Y, Z_{\hat{Y}}) = \sum_{j=1}^{p} B_j r_{yj} = \sum_{j=1}^{p} B_j^2 + 2\sum_{j=2}^{p}\sum_{i=1}^{p-1} B_i B_j r_{ij}$$

$$Var(Z_Y) = 1$$

$$Var(Z_{\hat{Y}}) = \sum_{j=1}^{p} B_j^2 + 2\sum_{j=2}^{p}\sum_{i=1}^{p-1} B_i B_j r_{ij}$$

$$R = Corr(Z_Y, Z_{\hat{Y}}) = \frac{\sum_{j=1}^{p} B_j^2 + 2\sum_{j=2}^{p}\sum_{i=1}^{p-1} B_i B_j r_{ij}}{\sqrt{\sum_{j=1}^{p} B_j^2 + 2\sum_{j=2}^{p}\sum_{i=1}^{p-1} B_i B_j r_{ij}}}$$

$$= \sqrt{\sum_{j=1}^{p} B_j^2 + 2\sum_{j=2}^{p}\sum_{i=1}^{p-1} B_i B_j r_{ij}} = \sqrt{\sum_{j=1}^{p} B_j r_{yj}}$$

where r_{yj} = dependent/independent variable Pearson (zero-order) correlations, and
r_{ij} = Pearson correlations among the *p* independent variables

Note: Proof of these formulas/relationships may be found in O'Brien (1982c).

As the reader can verify from Table 2, the covariance term is $Cov(Z_Y, Z_{\hat{Y}}) = Var(Z_{\hat{Y}}) = R^2$. These relationships constitute the "key" to the proof for the 0/1 limits of R as developed in this paper. We now demonstrate this proof. The development of the proof will consist of two parts --one part will demonstrate the proof for the lower limit and the other will show the proof for the upper limit. The lower limit is now presented.

Proof of the lower limit

The proof of the lower limit ($R \geq 0$) is based on an algebraic inequality and the information in Table 1 and Table 2. Recall the conceptual definition of the sample variance of $Z_{\hat{Y}}$:

$$Var(Z_{\hat{Y}}) = \frac{\sum_{i=1}^{n} (Z_{\hat{Y}_i} - \bar{Z}_{\hat{Y}})^2}{n-1}$$

As is true for any standardized mean, $\bar{Z}_{\hat{Y}} = 0$ (see Table 1). Thus,

$$Var(Z_{\hat{Y}}) = \frac{\sum_{i=1}^{n} Z_{\hat{Y}_i}^{\ 2}}{n-1}$$

The reader will agree that the following algebraic inequality is a true statement mathematically:

$$\frac{\sum_{i=1}^{r} {}'\,Y_i^{\ 2}}{n-1} \geq 0$$

From Table 2, this statement is equivalent to:

$$Var(\, Z_{\hat{Y}}) = \sum_{j=1}^{p} B_j^2 \; + \; 2 \sum_{j=2}^{p} \sum_{i=1}^{p-1} B_i B_j r_{ij}$$

But, as the reader can verify from Table 2, $Var(Z_{\hat{Y}}) = R^2$.

Hence,

$$\frac{\sum_{i=1}^{n} Z_{\hat{Y}_i}^{\;2}}{n-1} = Var(\, Z_{\hat{Y}}) = R^2 \quad \text{or}$$

$$Var(\, Z_{\hat{Y}}) \geq 0$$

Since the value of the square root of a variance term is, by definition, positive, then

$$\sqrt{Var(Z_{\hat{Y}})} \geq 0$$

or by substituting R^2,

$$\sqrt{R^2} \;\; \geq 0$$

Consequently,

$$R \geq 0.$$

The proof for the lower limit has been demonstrated.

Proof of the upper limit

The proof of the upper limit ($R \leq 1$) follows with similar logic. The reader will recall that the least squares criterion for standard scores can be stated as follows (see O'Brien, 1982c):

$$\sum_{i=1}^{n}(Z_{Y_i} - Z_{\hat{Y}_i})^2 = \text{ a minimum.}$$

We can also write the least squares criterion as:

$$\sum_{i=1}^{n}(Z_{Y_i} - Z_{\hat{Y}_i})^2 \geq 0$$

which is a true statement mathematically.

Our proof for the upper limit will consist of first expanding the above squared sum, substituting quantities from Tables 1 and 2, and simplifying. We then return to the inequality relation and conclude the derivation.

Expanding out the left side as a binomial and bringing in the summation operator:

$$\sum_{i=1}^{n}(Z_{Y_i} - Z_{\hat{Y}_i})^2 = \sum_{i=1}^{n} Z_{Y_i}^{\ 2} + \sum_{i=1}^{n} Z_{\hat{Y}_i}^{\ 2} - 2\sum_{i=1}^{n} Z_{Y_i} Z_{\hat{Y}_i}$$

Each term can be simplified in turn. As shown in Table 1, the sum of squared standardized scores in a sample is:

$\displaystyle\sum_{i=1}^{n} Z_{Y_i}^{\ 2} = n-1$ where n is the sample size. As for the second term in the expansion, that term reduces to

$$\sum_{i=1}^{n} Z_{\hat{Y}_i}^{2} = (n-1)Var(Z_{\hat{Y}})$$

which is derived as an algebraic manipulation for the form given in Table 1. The last term can be obtained in several steps by expansion and manipulation as follows:

$$2\sum_{i=1}^{n} Z_{Y_i} Z_{\hat{Y}_i} = 2\sum_{i=1}^{n} Z_{Y_i}(B_1 Z_1 + B_2 Z_2 + \ldots + B_p Z_p)$$

$$= 2\sum_{i=1}^{n}(B_1 Z_1 Z_{Y_i} + B_2 Z_2 Z_{Y_i} + \ldots + B_p Z_p Z_{Y_i})$$

$$= 2(B_1\sum_{i=1}^{n} Z_1 Z_{Y_i} + B_2\sum_{i=1}^{n} Z_2 Z_{Y_i} + \ldots + B_p\sum_{i=1}^{n} Z_p Z_{Y_i})$$

From Table 1, it can be seen that any term of the form $\sum_{i=1}^{n} Z_{X_i} Z_{Y_i}$ is equal to $(n-1)r_{xy}$. For correlations involving the independent/dependent variables (r_{yj}), we have:

$$2\sum_{i=1}^{n} Z_{Y_i} Z_{\hat{Y}_i} = 2\left[B_1(n-1)r_{y1} + B_2(n-1)r_{y2} + \ldots + B_p(n-1)r_{yp}\right].$$

$$= 2(n-1)\sum_{j=1}^{p} B_j r_{yj}.$$

Collecting all terms together, we can now rewrite the least squares criterion as:

$$\sum_{i=1}^{n}(Z_{Y_i} - Z_{\hat{Y}_i})^2 = n-1 + (n-1)Var(Z_{\hat{Y}}) - 2(n-1)\sum_{j=1}^{p}B_j r_{yj} \geq 0$$

Upon factoring out n-1 and dividing it through the inequality, we have

$$\sum_{i=1}^{n}(Z_{Y_i} - Z_{\hat{Y}_i})^2 = 1 + Var(Z_{\hat{Y}}) - 2\sum_{j=1}^{p}B_j r_{yj} \geq 0$$

Now, irom Table 2, we know that $\sum_{j=1}^{p}B_j r_{yj} = Var(Z_{\hat{Y}})$.

Thus,

$$\sum_{i=1}^{n}(Z_{Y_i} - Z_{\hat{Y}_i})^2 = 1 + Var(Z_{\hat{Y}}) - 2Var(Z_{\hat{Y}}) \geq 0$$

or

$$\sum_{i=1}^{n}(Z_{Y_i} - Z_{\hat{Y}_i})^2 = 1 - Var(Z_{\hat{Y}}) \geq 0$$

Reversing the sense of the inequality, we can write the right hand side of the above as:

$$Var(Z_{\hat{Y}}) \leq 1.$$

This gives equivalently,

$$\sqrt{Var(Z_{\hat{Y}})} \leq 1$$

But since $Var(Z_{\hat{Y}}) = R^2$, then

$\sqrt{R^2} \leq 1$ or

$R \leq 1$. Proof is completed.

We have proven in this paper that $0 \leq R \leq 1$.

References

Hays, W. L. *Statistics for the Social Sciences* (2nd ed.). NY: Holt, Rinehart & Winston, 1973.

Kendall, M. G. & A. Stuart. *The Advanced Theory of Statistics. Vol II: Inference and Relationship* (2nd ed.). NY: Hafner Publishing Co., 1967.

O'Brien, F. A proof that t^2 and F are identical: the general case. Washington, D.C. : Educational Resources Information Center, 1982a. (ED 215894)

O'Brien, F. Proof that sample bivariate correlation coefficient has limits ±. Washington, D.C. : Educational Resources Information Center, 1982b. (ED 216874)

O'Brien, F. A derivation of the sample multiple correlation formula for standard scores. Washington, D.C. : Educational Resources Information Center, 1982c. (ED 223429)

O'Brien, F. A derivation of the sample multiple correlation formula for raw scores. Washington, D.C. : Educational Resources Information Center, 1984. (ED 235205)

O'Brien, F. A derivation of the unbiased standard error of estimate: the general case. Washington, D.C. : Educational Resources Information Center 1987. (ED 280896)