THE SHRINKAGE OF THE COEFFICIENT OF MULTIPLE CORRELATION[1]

SELMER C. LARSON

Carleton College, Northfield, Minnesota

INTRODUCTION

It has been recognized by theoretical statisticians for some time that when the coefficient of multiple correlation (R) is derived for a given set of data, its value is likely to be deceptively large. If the computations have been correct, the value will hold rigidly for the set of data from which the regression equation was derived. If, however, the equation should be applied to a second set of data, even though strictly comparable, it has been supposed that the yield in this latter case would, except for errors due to sampling, be less than in the first. Moreover, it has been supposed that the more variables contained in the regression equation, the greater this shrinkage will be. This is particularly significant because ordinarily the practical employment of a regression equation involves its use with data other than those from which it was derived. If this shrinkage should turn out to be very large, the building of multiple regression equations might well be abandoned. The matter is therefore one of considerable importance, both theoretically and practically. Several attempts have been made by statisticians to derive a formula which should indicate the amount of this shrinkage. The most promising one of these will be considered in the present paper. So far as the writer has been able to discover, no one has attempted to determine experimentally the actual amount of shrinkage. The present report describes such an attempt in the field of psychological testing.

A study of the shrinkage is made by using a regression equation derived from one group of subjects to predict the criterion scores of a second group. The correlation yield by this procedure is subtracted from the yield obtained by predicting the criterion scores of the second group by means of a regression equation derived from themselves. This shrinkage is studied with the number of the independent variables in the regression equation ranging from one to ten in number and for a variety of different criteria. A comparison is then made between

[1] From the Psychological Laboratory, University of Wisconsin. The writer is greatly indebted to Professor M. V. O'Shea for permission to use data selected from the results obtained by the Mississippi Survey.

such empirical findings and the results obtained by applying a recently proposed formula for determining the same type of shrinkage.[1]

Source and Selection of Data

Something like 30,000 pupils were given mental and achievement tests in a survey of the school system of the State of Mississippi. For the present study, the test scores of eight hundred high school pupils from this number were used. The scores of pupils from the large and medium-sized high schools were chosen in the belief that the level of instruction would be more nearly uniform.[2] The entire population of eight hundred consisted of four groups—two hundred boys in each of two groups and two hundred girls in each of the two remaining groups. The subjects to make up these groups were chosen from those tested by the survey in such a way that each of the four contained exactly the same number of individuals drawn from any particular class of each school sampled. Otherwise the placing of the subjects in the several groups was entirely at random. By making up the personnel of the groups in this manner it was felt that they would be as exactly comparable in regard to general level and range of natural endowment, culture, and educational ooportunities as possible.

Each pupil had eighteen scores entered after his name. The designations are as follows:

X_1 English
X_2 Mathematics
X_3 Science
X_4 History
X_5 Chronological age
X_6 Intelligence quotient
X_7 Information (Terman)
X_8 Best answer (Terman)
X_9 Word meaning (Terman)
X_{10} Logical selections (Terman)
X_{11} Arithmetic (Terman)
X_{12} Sentence meaning (Terman)
X_{13} Analogies (Terman)
X_{14} Mixed sentences (Terman)
X_{15} Classifications (Terman)
X_{16} Number series (Terman)
X_{17} Total Terman
X_{18} Total Iowa

The first four X's together with X_{18} are scores made on the Iowa High School Content Examination. X_7 to X_{17} are scores made on the Terman Group Test of Mental Ability. Another column—the sum of each row—was added for checking purposes.

---

[1] Ezekiel, M J. B.: An unpublished paper read before the Mathematical Society at its annual meeting in Chicago in December, 1928.

[2] O'Shea's study showed that for the state as a whole scholastic achievement in the small high schools was decidedly lower than in the larger ones.

Empirical Determination of Shrinkage

Part I

Ten distinct regression equations with English as the criterion were derived from the data for the first group of boys. Ten parallel regression equations were derived from the data for the second group of boys. In the case of the first group, the first equation had all ten independent variables (tests). The second equation had the best nine test variables, *i.e.*, those having the highest criterion correlations. The third had the best eight test variables, and so on down to the tenth equation which was based on the single test having the highest criterion correlation. The same procedure was followed with the second group of boys, except that in this case the same test variables were used in the corresponding equations as were used with the first group. Owing to a natural variability in the size of the zero order correlation coefficients from sample to sample, the tests successively excluded from the progressively smaller equations with the second group of boys were not in all cases the next in order of weakness in the criterion r's.

Space is lacking for the presentation either of the means and standard deviations needed for the derivation of the regression equa-

Table I.—Zero Order Coefficients of Correlation for Both Sets of Boys

The Bold Face Figures at the Upper Right Are the Coefficients of Correlation for Boys, Set No. II; and the Light Face Figures at the Lower Left Are the Coefficients for Boys, Set No. I. At the Top and Left of Table Are Indicated the Variables Whose Notations Are Outlined Earlier in the Text. This Table Shows How English (X_1) Correlates with Each of the Items in the Terman Test and Also the Intercorrelations between the Various Items.

	x_1	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}
x_1	1 000	**.680**	**.533**	**.740**	**.531**	**.367**	**.502**	**.456**	**.418**	**.489**	**.204**	**.718**
x_7	699	1 000	**.714**	**.697**	**.631**	**.425**	**.565**	**.454**	**.451**	**.516**	**.264**	**.812**
x_8	499	617	1 000	**.602**	**.617**	**.447**	**.522**	**.433**	**.439**	**.442**	**.349**	**.786**
x_9	704	621	469	1 000	**.587**	**.348**	**.587**	**.446**	**.527**	**.531**	**.247**	**.837**
x_{10}	524	612	498	580	1 000	**.384**	**.466**	**.460**	**.391**	**.434**	**.286**	**.737**
x_{11}	366	316	272	392	382	1 000	**.280**	**.406**	**.210**	**.366**	**.416**	**.612**
x_{12}	506	477	414	567	407	382	1 000	**.297**	**.392**	**.356**	**.204**	**.688**
x_{13}	503	515	492	497	451	438	421	1 000	**.345**	**.362**	**.450**	**.653**
x_{14}	499	525	408	502	427	327	500	386	1 000	**.433**	**.254**	**.634**
x_{15}	474	516	400	505	470	335	390	499	359	1 000	**.288**	**.639**
x_{16}	255	316	334	350	424	553	342	534	280	418	1 000	**.545**
x_{17}	714	765	670	812	730	643	706	728	664	656	659	1 000

TABLE II.—SHOWING THE ACTUAL SHRINKAGE IN R FROM THE THEORETICAL VALUE OF AN EQUATION DERIVED FROM A GROUP OF SUBJECTS WHEN AN EQUATION DERIVED FROM A COMPARABLE GROUP IS APPLIED TO THEM. THE CRITERION THROUGHOUT IS HIGH SCHOOL ACHIEVEMENT IN ENGLISH

		Number of test variables used in prediction									
		1	2	3	4	5	6	7	8	9	10
Boys' Group No. 1.											
Correlation yield (R) from equations derived from their own scores	A	7042	7794	7834	7872	7880	7907	7929	7941	7944	7945
Correlation yield (R) from equations derived from scores of Group II	B	7042	7773	7798	7836	7820	7863	7827	7866	7847	.7832
Shrinkage	C	0000	0021	0036	0036	0060	0044	0102	0075	0097	0113
Boys' Group No. II.											
Correlation yield (R) from equations derived from their own scores	D	7402	7759	7813	7826	7847	7858	7859	7863	7868	7869
Correlation yield (R) from equations derived from scores of Group I	E	7402	.7728	7794	7803	7806	7725	7786	.7821	7794	.7786
Shrinkage	F	0000	0031	0019	0023	0041	0133	0073	0042	0074	.0083
Mean shrinkage of both groups $\frac{(C+F)}{2}$	G	0000	0026	.0027	0029	0050	0088	0087	0058	0085	0098

tions or for the regression equations themselves.[1] In order that the interested reader may study the relationship between the original correlations and the several multiple correlation yields, there has been placed in Table I the entire series of zero order correlation coefficients for both groups of boys. The coefficients of the respective groups are distinguished by means of contrasting type faces. All of the coefficients are positive.

The multiple correlation coefficients derived from the results shown in Table I are given in Table II. The R's corresponding to the several multiple regression equations derived from the boys of Group I are shown in row A and those for Group II in row D. These values are the correlation coefficients which would have been obtained if in each case the test scores from which the regression equation was derived had been substituted appropriately in the equation itself and the resulting criterion estimates had been correlated with the true criterion in the ordinary way. Actually, these values were obtained by means of the usual formula which is decidedly simpler. The coefficients in both series are distinctly high, as aptitude correlation yields run. It is noteworthy, however, that in both series alike, after three tests have been included in the equation, the addition of all the remaining seven tests suffices to raise the correlation yield a total of barely a single point in the second decimal place.

The next step in the process was to substitute the actual test scores of the boys of Group I in the equations derived from Group II and to correlate the resulting criterion estimates with the true criterion scores of Group I. The resulting series of coefficients is given in Table II, row B. The procedure was then reversed. The true criterion of Group II was correlated with the criterion estimates obtained by substituting their relevant test scores in the equations derived from the scores of Group I. The resulting coefficients are given in row E. We now have the values from which shrinkages may be determined.

According to the *a priori* expectation as indicated above, the values in row B should show a perceptible shrinkage when compared with the values in row A and similarly with row E when compared with row D. A brief comparison of the R values in the two pairs of rows shows that, except for the equation containing but a single test variable, this expectation is realized. The amount of the shrinkage is shown in

[1] These are given in detail for the entire study in the author's dissertation filed in the library of the University of Wisconsin. It is entitled "Studies in Aptitude Forecasting with the Multiple Regression Equation."

row *C* for the several pairs of values of Group I, and for Group II in row *F*. The mean values for rows *C* and *F* are given in row *G*. A glance at these latter values shows at once that, despite a certain amount of variability presumably due to sampling errors, there is a clear tendency for the shrinkage to increase with the increase in the number of independent (test) variables in the regression equation. This again is in harmony with what has been believed by statistical theorists. A graphic representation of the mean shrinkage values shown in row *G* is presented as the solid line in Fig. 1.

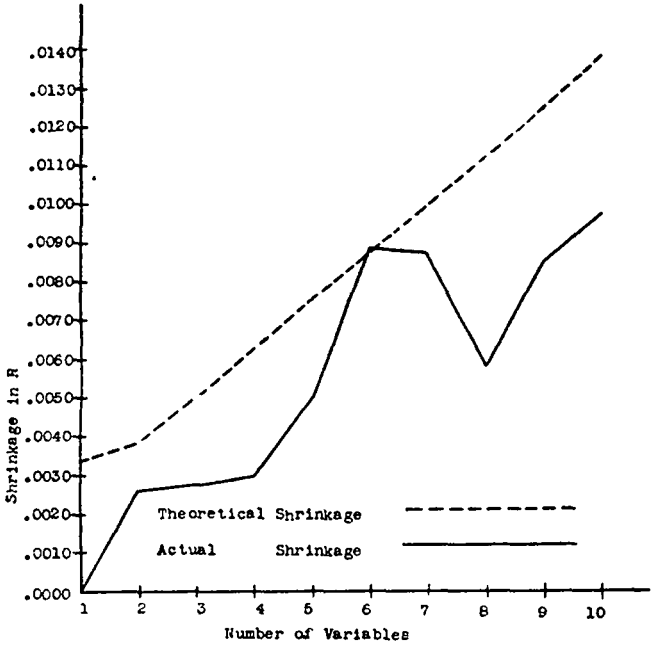

Fig. 1.—Shrinkage as obtained by the use of the formula and also as obtained experimentally.

We have seen that the increase in the size of *R*, even when regression equations are applied to the same data from which they are derived, is extremely slight and grows less and less as the number of test variables increases. We have also seen that under the same condition the amount of shrinkage in yield from such equations when in actual use grows greater and greater. The question naturally arises whether there may not be a point beyond which the increase in *R* resulting from the addition of a new test variable may not be more

than offset by the increase in shrinkage, so that the true functional value of such a test battery or other estimating aggregate may not be actually less than if the test or other independent variable had not been added. An examination of rows *B* and *E* shows that in the case of both groups alike such a critical point is reached at the eighth test added. *In both cases the batteries would have given an absolutely higher forecasting yield with two of the tests left out.* The moral of this is that under some circumstances the inclusion of certain tests in an aptitude battery after it has reached some size may not only entail the waste of energy and materials of administering the test, but may actually reduce the yield, and this even when the best possible method of weighting the tests is employed.

Part II

To secure further empirical evidence as to the amount of shrinkage from the ordinary multiple correlation coefficient, further computations

Table III.—Showing the Shrinkage of *R*'s When the Same (10) Test Variables Are Used but Different Academic Subjects Are Employed as Criteria with Different Correlation Yields

		English	Mathematics	Science	History	Total Iowa
Boys' Group II.						
Correlation yield (*R*) from equations derived from the subjects' own scores	*A*	7869	6431	5689	7719	8200
Correlation yield (*R*) from equations derived from the scores of Group I	*B*	7786	6148	.5230	7505	.8098
Shrinkage	*C*	0083	0283	0459	0214	.0102
Girls' Group II.						
Correlation yield (*R*) from equations derived from the subjects' own scores	*D*	7989	6403	4548	.7115	7875
Correlation yield (*R*) from equations derived from the scores of Group I	*E*	7665	6226	4219	6786	7755
Shrinkage	*F*	0324	0177	0329	0329	.0120

were undertaken in all of which the number of test variables was kept constant. The number chosen was the maximum for this study—ten. The multiple regression equations on mathematics, science,

history, etc., for the boys of Group I were used to estimate the corresponding true criterion scores of Group II. The same procedure was followed for the girls.

The results are given in Table III which is constructed in a manner comparable to Table II above. With the number of test variables constant, this table enables us to observe the influence upon the amount of shrinkage of the strength of the natural tendency to correlation in the test data involved. If we divide the ten shrinkages found in this series into two groups on the basis of the size of the original R's, we find that the average shrinkage for the five largest R's is .0169 whereas that for the five smallest R's is .0315. The tendency for the weaker sets of data to yield the larger shrinkages is evident. For the two lowest values (Science) this amounts to .0394, a very appreciable amount.

The Smith Shrinkage-deduction Formula

A promising correction formula has been developed to apply to the coefficient of multiple correlation. A paper containing the formula was read by M. J. B. Ezekiel at the December 1928 meeting of the American Mathematical Society held at Chicago. This formula is

$$\bar{R}^2 = 1 - \frac{1 - R^2}{1 - \frac{m}{n}} \quad \text{or} \quad \bar{R}^2 = \frac{nR^2 - m}{n - m}$$

where $\bar{R}$ = the estimated correlation obtaining in the universe
R = the observed correlation
m = the number of variables, dependent and independent
n = the number of observations (statistical population)

Ezekiel gives the credit for developing this formula to B. B. Smith.

At the completion of the computations described in the preceding section it was a relatively simple task to substitute in the above formula and determine for the various observed R's, the corresponding estimates of the "correlation obtaining in the universe." These values are given in Table IV for the R's obtained in Part I. Table V shows the corresponding values of the R's obtained in Part II. Subtraction from the original R's shown in Tables II and III respectively yields the amounts of shrinkage which would be anticipated by the formula in each case.

In Table IV row E shows the mean amount of shrinkage for each pair of observed R values for the several numbers of test variables.

TABLE IV.—SHOWING THE SHRINKAGE OF R'S AS INDICATED BY THE SMITH FORMULA, WHERE THE VARIABLES RANGE IN NUMBER FROM ONE TO TEN

		Number of test variables used in prediction									
		1	2	3	4	5	6	7	8	9	10
Boys' Group I.											
Correlation obtaining in the universe (R) as estimated by the Smith formula	A	7006	.7756	7784	7810	.7805	.7821	.7831	.7831	7821	.7809
Shrinkage obtained by subtracting the values in row A above from those in row A of Table II	B	0036	0038	.0050	0062	0075	.0086	0098	0110	.0123	0136
Boys' Group II.											
Correlation obtaining in the universe ($\overline{R}$) as estimated by the Smith formula	C	7371	.7720	.7762	7762	7771	.7769	7757	.7748	7740	.7727
Shrinkage obtained by subtracting the values in row C above from those in row D of Table II	D	0031	0039	0051	.0064	.0076	.0089	.0102	0115	0128	.0142
Mean shrinkage of both groups $\frac{(B+D)}{2}$	E	.0034	0039	.0051	0063	0076	0088	.0100	.0113	.0126	.0139

It is plotted in Fig. 1 for the purpose of easy comparison with the empirically determined shrinkages. It is clear that the formula definitely parallels the empirical findings. It conforms to the observed tendency for the amount of shrinkage to increase with the number of variables involved in the equation. There is a well-marked tendency, however, for the Smith formula to indicate materially larger shrinkages than the empirical results show.

Passing to Table V, we may make the comparison with the empirical results by treating the shrinkages the same as those in Table III.

TABLE V.—SHOWING THE SHRINKAGE OF R'S AS INDICATED BY THE SMITH FORMULA, WHERE THE NUMBER OF VARIABLES IS CONSTANT BUT THE SIZE OF THE R'S VARY RATHER WIDELY

		English	Mathematics	Science	History	Total Iowa
Boys' Group II.						
Correlation obtaining in the universe ($\bar{R}$) as estimated by the Smith formula	*C*	7727	.6156	5330	7563	8094
Shrinkage obtained by subtracting row *A* above from row *A* in Table III.	*B*	0142	.0275	0359	0156	.0106
Girls' Group II.						
Correlation obtaining in the universe ($\bar{R}$) as estimated by the Smith formula	*C*	7843	.6132	.4005	.6916	7730
Shrinkage obtained by subtracing row *C* above from row *D* in Table III.	*D*	0146	0271	0543	0199	0145

Computation shows that the mean shrinkage of the five largest R's is .0139 whereas that for the five smallest is .0329. Here again we observe, as in Table IV, that the shrinkages yielded by the formula are materially larger than those found empirically. An analysis of the formula reveals that the shrinkage increases as the size of the obtained R decreases. The formula will break down, however, when $m/n > R^2$ as the values will then become imaginary. In this situation, with m equal to 11 and n equal to 200, the formula will give imaginary values when the absolute values of R are less than .235.

Summary and Conclusions

1. The present investigation has shown that the theoretically expected shrinkage of R as derived by the multiple correlation formula is a fact.

2. The shrinkage is found to increase as the number of test variables increases.

3. The shrinkage is also found to increase as the size of R decreases.

4. The Smith shrinkage-deduction formula parallels all of the above empirical findings, but quite consistently gives values which are in excess of those obtained under the present experimental conditions.

5. The empirically observed shrinkage increases at such a rate with the increase in the number of test variables that one of the most widely known scholastic aptitude tests actually shows a lower correlation yield with a criterion when ten test units are used than when only eight are employed. This suggests that test batteries may have very definite limitations as to size.