DOCUMENT RESUME

ED 262 081 TM 850 562

AUTHOR Wolfle, Lee M.; Ethington, Corinna A.
TITLE Robustness of Regression Estimates for Ordered Dichotomous Variables.
PUB DATE Mar 85
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Correlation; Estimation (Mathematics); Matrices; *Multiple Regression Analysis; Simulation; Statistical Studies; *Validity
IDENTIFIERS Dichotomous Variables; *Pearson Product Moment Correlation; Polyserial Correlation; *Robustness; Tetrachoric Correlation

ABSTRACT

The purpose of this paper is to examine the validity of regression estimates when skewed dichotomous scales are used as independent variables. When Pearson product-moment correlations are used to measure zero-order associations involving dichotomous variables, the resulting coefficients underestimate the true associations. As a result, using product-moment correlations involving dichotomous variables in regression equations apparently yields biased partial regression estimates. The analysis reported here was based on fifty sets of simulated data with 500 cases and four independent variables in each set. We found that tetrachoric and polyserial correlations for associations involving dichotomized variables yield more accurate estimates of the regression coefficients based on the underlying continuous data than did product-moment correlations for the same dichotomous variables. (Author)

ROBUSTNESS OF REGRESSION ESTIMATES FOR ORDERED DICHOTOMOUS VARIABLES

Lee M. Wolfle

Virginia Polytechnic Institute and State University
and
Educational Testing Service

and

Corinna A. Ethington

Virginia Polytechnic Institute and State University

This paper was prepared for delivery at the annual meetings of the American Educational Research Association, Chicago, March 31 - April 4, 1985.

ROBUSTNESS OF REGRESSION ESTIMATES FOR ORDERED DICHOTOMOUS VARIABLES

ABSTRACT

The purpose of this paper is to examine the validity of regression estimates when skewed dichotomous scales are used as independent variables. When Pearson product-moment correlations are used to measure zero-order associations involving dichotomous variables, the resulting coefficients underestimate the true associations. As a result, using product-moment correlations involving dichotomous variables in regression equations apparently yields biased partial regression estimates. The analysis reported here was based on fifty sets of simulated data with 500 cases and four independent variables in each set. We found that tetrachoric and polyserial correlations for associations involving dichotomized variables yield more accurate estimates of the regression coefficients based on the underlying continuous data than did product-moment correlations for the same dichotomous variables.

ROBUSTNESS OF REGRESSION ESTIMATES FOR ORDERED DICHOTOMOUS VARIABLES

One of the most frequently used measures of association is the product-moment correlation coefficient. Not only are correlation coefficients used by themselves as bivariate measures of association, they are frequently used in matrix form as input into multivariate analyses such as factor analysis or multiple regression. While no assumptions are required for the computation of a product-moment correlation, the interpretation of the resulting coefficient certainly depends on whether or not the data fit an appropriate statistical model (Carroll, 1961). In particular, the product-moment correlation measures a linear relationship between two continuous variables. When a product-moment correlation is used to estimate the degree of association between two categorical variables with underlying continuities, the possible range of values for a product-moment correlation may not be from -1.0 to +1.0, but is dependent on the marginal distributions. The true association between two categorical variables will be underestimated with a product-moment correlation (Carroll, 1961; Ferguson, 1941; Muthen, 1983a, 1983b; Pearson, 1900, 1904, 1913).

Muthen (1983a) vividly illustrated the underestimation of the association between two categorical variables when estimated with product-moment correlations. Muthen began his illustration with two continuous variables whose true correlation was .50. He then categorized the same data into two, three, four, and five categories with varying degrees of skewness. Such variables are those one would

encounter, for example, if one used an agree-disagree questionnaire format or a Likert scale to measure some underlying continuous attitudinal scale. For two dichotomous variables with zero skewness (i.e., a 50/50 split) Muthen showed that the product-moment correlation was .33, and decreased to a mere .10 when a variable split 90/10 was correlated with a variable split 10/90. Thus, the degrees of association among all categorical variables are underestimated with product-moment correlations, but the greatest underestimation occurs with highly skewed dichotomous variables.

Associations among categorical variables are more appropriately measured by tetrachoric or polychoric correlations (Carroll, 1961; Muthen 1983a, 1983b; Pearson, 1913; Pearson and Pearson, 1922). Such correlations are estimates of population correlations among latent, continuous response variables (Brown and Benedetti, 1977; Muthen, 1983b), and are calculated with reference to threshold values estimated from the marginal distributions. The use of these correlations may be thought of as robustifying the correlations against categorization and skewness, or "stretching" the correlations to assume values between -1.0 and +1.0 (Muthen, 1983b).

A related instance of an underestimation of association occurs when product-moment correlations are used to measure the degree of association between continuous and ordered categorical variables. Such associations are more appropriately measured with polyserial correlations (Jaspen, 1946; Olsson, Drasgow, and Dorans, 1982; Pearson, 1913). As before, the degree of underestimation is greatest when one of the variables is a dichotomy and both variables are highly skewed.

For a long time it has been known that product-moment correlations underestimate associations among ordered categorical data. When product-moment correlations are used to measure associations among such data, and then used in matrix form as input into multivariate analyses such as factor analysis or multiple regression, the consequences are less clear. Muthen (1983a) and Olsson (1979) have demonstrated in factor analytic models that the use of tetrachoric in place of product-moment correlations produced more satisfactory solutions. Using product-moment correlations and standard estimation procedures in such situations resulted in downwardly biased estimates of the factor loadings of the categorical variables, and also frequently resulted in a greater number of factors.

The question immediately arises whether multiple regression estimates will also be affected by the choice of zero-order measures of association for ordered categorical variables. Consider, for example, an equation with two independent variables, in which the dependent variable and the first independent variable are measured on continuous scales, while the second independent variable is a skewed dichotomous scale with a 70/30 split. Furthermore, consider that the population correlations among all three variables are .50, .50, and .50. The resulting standardized regression equation would be:

$$Y' = .33X_1 + .33X_2.$$

If product-moment correlations had been used to measure the zero-order associations for all three variables, the resulting correlation coefficients would have been .50, .38, and .38, respectively (Muthen, 1983a), and

the resulting standardized regression equation would be:

$$Y' = .42X_1 + .22X_2.$$

Using product-moment correlation coefficients would thus underestimate the true influence of the dichotomous variable (X_2) on the dependent variable. Furthermore, the regression coefficient for the continuous independent variable (X_1) would overestimate its true effect, since the estimated correlation between the two independent variables is less than the true level of their association.

While appropriate estimation procedures for measuring associations involving dichotomous (or other ordered polychotomous) independent variables have been known for some time, the complexity of the calculations involved has prevented their use in practice. New developments, however, now make their calculation relatively easy (Joreskog and Sorbom, 1983; Muthen, 1982). Even so, such associations are still often measured with product-moment correlation coefficients. Such zero-order correlations, however, are known to be biased, and apparently yield biased partial regression estimates. Accordingly, the purpose of this paper is to examine the robustness of regression estimates when skewed dichotomous variables are included among the independent regressors.

METHODS

There are, of course, a wide variety of models that could have been estimated in this study. Here we chose only one. The model used here contained four independent variables. Two of the variables

(X_1 and X_2) may be considered, say, social background variables measured on continuous scales, and only weakly or modestly related to the dependent variable, which is often the case in reality. The other two variables (X_3 and X_4) are more strongly related to the dependent variable and to each other.

The assessment of the robustness of regression coefficients estimated from product-moment correlation coefficients for dichotomous variables was based on the analysis of fifty sets of simulated data. Each of these sets contained 500 cases with data for five standard normal variables. The averages of the product-moment correlations among these continuous variables are shown in Table 1.

Insert Table 1 About Here

Using the continuous variables, the regression equation was estimated with each of the fifty sets of data. For each independent variable, the fifty estimated regression coefficients were then averaged. The averages for the continuous-data regression coefficients represent the most accurate estimates of the influence of the four independent variables on the dependent variable. Accordingly, these averages were the standard against which subsequent estimates were compared in order to determine the effect of categorization and correlation type on regression estimates.

After estimating the regression equations with continuous data, two independent variables (X_3 and X_4) were dichotomized in three

different patterns: both were split 50/50, both were split 80/20, and the first one was split 80/20 while the second was split 20/80. The zero-order associations among these variables, for each of the various methods of dichotomizing the variables, were first estimated with product-moment correlations and the regression equations estimated with SPSS^X (SPSS, Inc., 1983). The zero-order associations were then re-estimated with a mixed matrix of product-moment, tetrachoric, and polyserial correlations, as appropriate, using LISREL-VI (Joreskog and Sorbom, 1983), and the regression equations estimated with SPSS^X. (Note that generating the data as standard normal variables guarantees that subsequent dichotomization and the use of tetrachoric and polyserial correlations meets the assumption that continuous distributions underlie the dichotomized variables.)

To determine the robustness of the regression estimates, the fifty sets of regression estimates were averaged for each of the six methods used to estimate the regression equations containing the dichotomized variables (i.e., 50/50 with product-moment and with mixed correlations; 80/20 with product-moment and with mixed correlations; and 80/20 - 20/80 with product-moment and with mixed correlations). Comparisons of these averages to the average regression coefficients based on the continuous data allow us to determine the impact of dichotomization on the assessment of the influence of the independent variables on the dependent variable. Those averages closely approximating their corresponding regression coefficients based on the continuous data were considered robust estimates. Accurate average estimates, however, are

of little value if they exhibit large variability. Accordingly, the variances of the estimates were also computed, and are reported below.

We hypothesized that using tetrachoric and polyserial correlations for associations involving dichotomous variables would yield more accurate estimates of the regression effects than would product-moment correlations for the same variables. Furthermore, we hypothesized that the greater the degree of skewness among the independent variables, the greater would be the bias among regression estimates based on product-moment correlations.

RESULTS

In the regression results reported below there are four independent variables. Two of them (X_1 and X_2) were generated as continuous standard normal variables, and were used as such in all of the regression runs. Two others (X_3 and X_4) were generated as continuous standard normal variables, but subsequently were split into dichotomies as discussed above.

Table 2 shows the results of these regressions. The first row of regression coefficients in Table 2 contains the average of the fifty estimations of the regression equation when all four variables were entered as continuous data. These coefficients, therefore, are the standards against which all other results should be compared. The next three rows of regression coefficients are the averages of the fifty estimations of the regression equation for the various splits; the input correlation matrix consisted entirely of product-moment correlations. The

last three rows of regression coefficients in Table 2 are the averages for the various splits; in these cases the input correlation matrix consisted of a mixture of product-moment, tetrachoric, and polyserial correlations, as appropriate.

Insert Table 2 About Here

Examination of the coefficients shown in Table 2 clearly indicates that the use of tetrachoric and polyserial correlations for dichotomized variables provide better average estimates of the effects of the independent variables than did the use of product-moment correlations. The average regression coefficient for X_1 when all four independent variables were continuous was -.085. For each degree of skewness, the average regression coefficients for X_1 using tetrachoric and polyserial correlations were very close to -.085; none of the coefficients based on product-moment correlations was as close. The explanation is not complicated. Both Y, the dependent variable, and X_1 were measured as continuous variables; their average zero-order association, however, was small (.067). But X_1 was also correlated with the other independent variables, and when these other variables were controlled the partial effect of X_1 became negative. With product-moment correlations, however, the zero-order associations between X_1 and the two dichotomized variables (X_3 and X_4) were underestimated; thus, the influence of X_1 on Y was an attenuated estimate. These correlations, for both the product-moment and the mixed matrices, can

be found in Tables 3, 4, and 5, respectively, for the three different splits.

Insert Tables 3, 4, and 5 About Here

The same conclusions can be seen to apply also to the regression coefficients for X_2. The use of tetrachoric and polyserial correlations led to regression estimates much closer to the regression coefficient based on continuous data than did the use of product-moment correlations. Once again, the product-moment correlations underestimate the associations between X_2 and X_3, and between X_2 and X_4; thus, the partial effect of X_2 on Y was overestimated.

In contrast, the use of product-moment correlations led to an underestimation of the partial effect of X_3 on Y. In this case, X_3 was a dichotomized variable, and the use of a product-moment correlation underestimated its zero-order association with the dependent variable. Of course, the associations among X_3 and the other independent variables were also underestimated with product-moment correlations; but correcting these underestimated associations (particularly with X_1 and X_2) by using polyserial and tetrachoric correlations had only a modest influence on the estimated partial effect of X_3 on Y in comparison with correcting the estimated association between X_3 and Y itself.

Finally, examination of the regression coefficients for X_4 shown in Table 2 indicates no systematic bias in the use of product-moment versus polyserial and tetrachoric correlations. In this case, the use of

product-moment correlations not only underestimated the association between X_4 and Y, but also underestimated the rather substantial correlations between X_4 and the other independent variables. Consequently, using product-moment correlations to control for X_1, X_2, and X_3 in measuring the partial influence of X_4 on Y removed less of an effect than the statistical control should have, but removed it from an attenuated estimate of the association between X_4 and Y. In contrast, using polyserial and tetrachoric correlations removed more of an effect of X_4 on Y, but removed it from a disattenuated estimate of the zero-order association between X_4 and Y.

In summary, the estimated partial regression coefficients for continuous independent variables were more accurately estimated on the average by using a mixed matrix of tetrachoric, polyserial, and product-moment correlations when some of the variables were dichotomized than by using product-moment correlations alone, which underestimated zero-order associations involving categorical variables.

We also hypothesized that the greater the degree of skewness among the independent variables, the greater would be the bias among the estimates based on product-moment correlations. This hypothesis seems to be false. An examination of the average coefficients within correlation type in Table 2 shows that they differ only slightly. Furthermore, it appears that the average coefficients for degree of split could have varied by chance alone. To test this, a series of two-way analyses of variance were performed for each set of regression coefficients. While significant differences (with $\alpha = .01$) were found

between correlation-matrix type for X_1, X_2, and X_3 (but not for X_4), no significant differences were found among the three degrees of skewness for any of the four independent variables.

CONCLUSIONS

While it has long been known that Pearson product-moment correlations underestimate zero-order associations involving ordered categorical variables, effects on multivariate partial estimates have only recently been investigated. Muthen (1983a) and Olsson (1979) concluded that the factor loadings in factor analyses involving dichotomized variables were better estimated with tetrachoric than with product-moment correlations.

This paper investigated the effects on partial regression coefficients of using an input matrix of product-moment correlations versus a mixed matrix of tetrachoric, polyserial, and product-moment correlations. We concluded that the average of fifty replications using simulated data with a mixed matrix provided more accurate estimates of the regression coefficients based on continuous data than were the estimates generated from the input of a matrix of product-moment correlations.

This does not mean, however, that the use of tetrachoric and polyserial correlations is cost free. In the first place, it is possible for a mixed matrix of tetrachoric, polychoric, polyserial, and product-moment correlations not to be positive definite (Joreskog and Sorbom, 1983, p. IV.6). Thus, there would be no regression solution.

Even when the problem of positive definiteness does not develop (and never occurred in the 350 replicated regressions reported here) other costs develop. In particular, the variance of the estimates computed from tetrachoric and polyserial correlations were greater in value than the estimates generated from product-moment correlations. This is shown in Table 6, where one can see that every standard deviation for the fifty coefficients for product-moment correlations was less than the corresponding standard deviation for the coefficients using tetrachoric and polyserial correlations.

Insert Table 6 About Here

Thus, one must balance the costs and benefits of using as input into regression analyses product-moment versus tetrachoric and polyserial correlations for ordered dichotomous variables. The conservative choice is to use product-moment correlations, recognizing that the true effects may be underestimated. The use of tetrachoric and polyserial correlations on the average produce more accurate regression parameter estimates, but the greater variability of the estimates can, for a single case, result in estimates far from the true parameter. Knowing this, it would be advisable to cross validate one's results when using these correlations.

REFERENCES

Brown, M.B., & Benedetti, J.K. (1977). On the mean and variance of the tetrachoric correlation coefficient. Psychometrika, 42, 347-355.

Carroll, J.B. (1961). The nature of the data, or how to choose a correlation coefficient. Psychometrika, 26, 347-372.

Ferguson, G.A. (1941). The factorial interpretation of test difficulty. Psychometrika, 6, 323-329.

Jaspen, N. (1946). Serial correlation. Psychometrika, 11, 23-30.

Joreskog, K.G., & Sorbom, D. (1983). LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Squares Methods. Chicago: National Educational Resources.

Muthen, B. (1982). LACCI: Latent variable analysis with dichotomous, ordered categorical, and continuous indicators - An experimental program for researchers. Unpublished.

Muthen, B. (1983a, April). Categorical vs. continuous variables in factor analysis and structural equation modeling. Paper presented at the meeting of the American Educational Research Association, Montreal.

Muthen, B. (1983b). Latent variable structural equation modeling with categorical data. Journal of Econometrics, 22, 43-65.

Olsson, U. (1979). On the robustness of factor analysis against crude classification of the observations. Multivariate Behavioral Research, 14, 485-500.

Olsson, U., Drasgow, F., & Dorans, N. (1982). The polyserial correlation coefficient. Psychometrika, 47, 337-347.

Pearson, K. (1900). Mathematical contributions to the theory of evolution, VII: On the correlation of characters not quantitatively measurable. Philosophical Transactions of the Royal Society of London A, 195, 1-147.

Pearson, K. (1904). Mathematical contributions to the theory of evolution, XIII: On the theory of contingency and its relation to association and normal correlation. Drapers Company Research Memoirs, Biometric Series, No. 1.

Pearson, K. (1913). On the measurement of the influence of 'broad categories' on correlation. Biometrika, 9, 116-139.

Pearson, K., and Pearson, E.S. (1922). On polychoric coefficients of correlation. Biometrika, 14, 127-156.

SPSS, Inc. (1983). SPSS-X User's Guide. New York: McGraw-Hill Book Co.

Table 1. Average Product-Moment Correlations among Five Standard Normal Continuous Variables

	Y	X_1	X_2	X_3	X_4
Y	---				
X_1	.067	---			
X_2	.204	.344	---		
X_3	.547	.128	.120	---	
X_4	.479	.261	.323	.553	---

Table 2. Summary of Regression Results

Correlation Type/ Split	Independent Variables				R^2
	X_1	X_2	X_3	X_4	
Pearson/ Continuous	-.0848	.1069	.4136	.2369	.3600
Pearson/ 50/50	-.0612	.1304	.3404	.2472	.2661
80/20	-.0487	.1474	.3033	.2124	.2195
80/20-20/80	-.0556	.1426	.3238	.2451	.2357
Tetrachoric etc./					
50/50	-.0834	.1016	.4092	.2473	.3632
80/20	-.0832	.1046	.4112	.2450	.3654
80/20-20/80	-.0859	.1078	.4138	.2405	.3640

Table 3. Average Correlations for 50/50 Split[a]

	Y	X_1	X_2	X_3	X_4
Y	---	.067	.204	.434	.382
X_1	.067	---	.344	.098	.210
X_2	.204	.344	---	.101	.255
X_3	.545	.123	.127	---	.366
X_4	.480	.263	.320	.544	---

[a]Product-moment matrix is above the diagonal and the mixed matrix is below.

Table 4. Average Correlations for 80/20 Split[a]

	Y	X_1	X_2	X_3	X_4
Y	---	.067	.204	.381	.339
X_1	.067	---	.344	.088	.079
X_2	.204	.344	---	.079	.234
X_3	.544	.125	.113	---	.333
X_4	.485	.255	.322	.550	---

[a]Product-moment matrix is above the diagonal and the mixed matrix is below.

Table 5. Average Correlations for 80/20-20/80 Split[a]

	Y	X_1	X_2	X_3	X_4
Y	---	.067	.204	.381	.334
X_1	.067	---	.344	.088	.183
X_2	.204	.344	---	.079	.223
X_3	.544	.125	.113	---	.207
X_4	.477	.266	.322	.546	---

[a]Product-moment matrix is above the diagonal and the mixed matrix is below.

Table 6. Standard Deviations for Average Regression Coefficients Shown in Table 1

Correlation Type/ Split	Independent Variables				R^2
	X_1	X_2	X_3	X_4	
Pearson/ Continuous	.0403	.0363	.0372	.0420	.0320
Pearson/ 50/50	.0441	.0386	.0325	.0384	.0325
80/20	.0462	.0406	.0297	.0374	.0296
80/20-20/80	.0415	.0382	.0363	.0345	.0292
Tetrachoric etc./					
50/50	.0472	.0427	.0538	.0613	.0436
80/20	.0508	.0484	.0557	.0708	.0467
80/20-20/80	.0421	.0430	.0726	.0754	.0464