The Correction for Restriction of Range and Nonlinear Regressions: An Analytic Study

Alan L. Gross and Lynn E. Fleischman
City University of New York

The effect of a nonlinear regression function on the accuracy of the restriction of range correction formula was investigated using analytic methods. Expressions were derived for the expected mean square error (EMSE) of both the correction formula and the squared correlation computed in the selected group, with respect to their use as estimators of the population relationship. The relative accuracy of these two estimators was then studied as a function of the form of the regression, the form of the marginal distribution of x scores, the strength of the relationship, sample size, and the degree of selection. Although the relative accuracy of the correction formula was comparable for both linear and concave regression forms, the correction formula performed poorly when the regression form was convex. Further, even when the regression is linear or concave, it may not be advantageous to employ the correction formula unless the xy relationship is strong and sample size is large.

In validating a test x as a predictor of some criterion y, an investigator must often deal with the restriction of range problem, that is, the problem of missing data on the criterion variable. Typically, the missing data are the result of a selection process which eliminates persons before y scores can be observed. The traditional solution to the problem is to employ the so-called correction formula (Lord & Novick, 1968). Given the xy data of the selected persons and the x scores of the unselected persons,
the formula provides an estimate of the xy correlation in the total group.

An important assumption underlying the correction formula is the linearity of the regression of y on x. In other words, it is assumed that the regression line computed in the selected group can be extrapolated to the full sample. The important question of how robust the correction formula estimate is to violations in the linearity assumption has been investigated empirically by various researchers (Greener & Osburn, 1979, 1980; Gross, 1982; Gross & Fleischman, 1983). The present study extended this research by using analytic methods to investigate the effect of nonlinearity on the accuracy of the correction formula.

The problem of a nonlinear regression form becomes especially difficult when data are missing on the criterion variable. If complete xy data could be observed for all persons, the regression function could be linearized by employing a suitable transformation. However, when data are missing on y as a function of selection on x (e.g., when the persons scoring highest on x are selected and observed on y), the nonlinearity of the data may not be apparent to the investigator. In other words, the regression function may appear to be linear in the range of the observed data. Consequently, the correct transformation of the data may be difficult to obtain. Because an investigator may not always be able to ascertain whether the regression function is linear, it is important to consider the question of

the robustness of the correction formula to a non-linear regression function. Such functions may occur in psychological studies where ceiling or floor effects are operating.

The basic approach in the present study is to consider a total population where the regression of y on x is nonlinear. An expression is then derived for the expected mean square error (EMSE) of the correction formula (r^2_{xyc}) as an estimate of the population xy relationship. The accuracy of the squared xy correlation computed in the selected group (r^2_{xys}) is investigated by deriving an expression for the EMSE of this statistic as an estimate of the population relationship. The accuracy (i.e., the EMSE) of the two estimates is then compared as a function of the following factors: (1) the form of the marginal distribution of x scores, (2) the strength of the population xy relationship, (3) sample size, and (4) the degree of selection. The results of the investigation permit definition of the conditions under which it is advantageous to use the correction formula as an estimate of the population relationship, rather than the squared xy correlation computed in the selected group.

Method

EMSE of the Estimators

First consider the population relationship between the x and y variables. Given x, suppose that y is normally distributed with mean

$$E(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{1}$$

and conditional variance σ_y^2. The xy relationship can be said to be nonlinear in the population when the β_2 term is not zero, otherwise it is linear. Now consider a sample consisting of N xy pairs drawn from the population. The linear xy relationship in the sample can be described by the squared product moment correlation:

$$r^2_{xy} = s^2_{xy}/(s^2_x s^2_y) \quad , \tag{2}$$

where s_{xy}, s^2_x, and s^2_y are the xy covariance, x variance, and y variance, all computed in the total sample. The expected value of this squared correlation over all total samples of size N, $E(r^2_{xy})$, defines a parameter which can be taken to represent

the population relationship between x and y. More specifically, $E(r^2_{xy})$ measures the "average" validity of x as a linear predictor of y in a total sample of size N.

If there were no missing data on the y variable, the population relationship $E(r^2_{xy})$ could be simply estimated by computing the squared xy correlation given in Equation 2 using the complete set of x and y scores observed in a sample of size N. However, when a selection process is operating, this estimate cannot be computed due to the missing y scores of the unselected persons. In this case, the correction formula (r^2_{xyc}) and the squared xy correlation computed in the selected sample (r^2_{xys}) can be used as estimates of the $E(r^2_{xy})$ parameter. These two estimates are expressed as

$$r^2_{xyc} = \frac{r^2_{xys}}{[r^2_{xys} + (s^2_{xs}/s^2_x)(1 - r^2_{xys})]} \tag{3}$$

and

$$r^2_{xys} = \frac{s^2_{xys}}{s^2_{xs} s^2_{ys}} \tag{4}$$

where s_{xys}, s^2_{xs}, s^2_{ys} are the xy covariance, variance of x, and variance of y in the selected group, and s^2_x is the variance of x in the total group.

The accuracy of the estimators expressed in Equations 3 and 4 can be assessed in terms of the criterion of expected mean square error. The EMSE for the correction formula is expressed as

$$\begin{aligned} \text{EMSE}(r^2_{xyc}) &= E[r^2_{xyc} - E(r^2_{xy})]^2 \\ &= \text{Var}(r^2_{xyc}) + [E(r^2_{xyc}) - E(r^2_{xy})]^2 \\ &= \text{Variance} + \text{Bias}^2 \end{aligned} \tag{5}$$

Similarly, the EMSE for the squared correlation computed in the selected group is given as

$$\begin{aligned} \text{EMSE}(r^2_{xys}) &= E[r^2_{xys} - E(r^2_{xy})]^2 \\ &= \text{Var}(r^2_{xys}) + [E(r^2_{xys}) - E(r^2_{xy})]^2 \\ &= \text{Variance} + \text{Bias}^2 \end{aligned} \tag{6}$$

In order to evaluate the EMSE measures given in Equations 5 and 6, expressions must be derived for the following parameters: $E(r^2_{xy})$, $E(r^2_{xyc})$, $E(r^2_{xys})$, $\text{Var}(r^2_{xyc})$, and $\text{Var}(r^2_{xys})$. It should be noted that the three squared correlations, r^2_{xy}, r^2_{xyc}, and r^2_{xys}, can be expressed as simple functions of a random variable having a doubly noncentral F distribution (Searle, 1971). Thus, the statistical problem of deriving

expressions for the expected values and variances is basically one of computing the expected value and variance of functions of a doubly noncentral F distribution.

The general approach can be illustrated by first considering the derivation of expressions for the expected value and variance of the squared xy correlation computed in the total group. This statistic can be expressed in terms of the following function of the doubly noncentral F'' variable:

$$r_{xy}^2 = F''/(N-2+F'') \quad , \tag{7}$$

where

$$F'' = (N-2)(SS_{\hat{y}}/SS_e) \tag{8}$$

and $SS_{\hat{y}}$ and SS_e are the sum of squares of predicted y and the sum of squares of error in the total sample, respectively.

The F'' variable in Equation 7 has a doubly noncentral F distribution with degrees of freedom 1 and $N-2$, and noncentrality parameters λ_1 and λ_2 expressed as follows:

$$\lambda_1 = \frac{\beta'X'[Z(Z'Z)^{-1}Z' - E/N]X\beta}{2\sigma_y^2} \tag{9}$$

$$\lambda_2 = \frac{\beta'X'[I - Z(Z'Z)^{-1}Z']X\beta}{2\sigma_y^2} \tag{10}$$

where β is a 3×1 vector of population regression weights;

X is an $N \times 3$ data matrix consisting of a leading column of unities, a second column of x scores, and a third column of x^2 scores;

Z represents the first two columns of the X matrix;

E is an $N \times N$ matrix of unities; and

I is an $N \times N$ identity matrix.

It should be noted that if the population regression is linear (i.e., $\beta_2 = 0$), then $\lambda_2 = 0$, and the F'' variable in Equation 7 has a singly noncentral distribution. Further, if there is no relationship between x and y in the population (i.e., both β_1 and β_2 are 0), then $\lambda_1 = \lambda_2 = 0$, and the F'' variable has an ordinary central F distribution.

The expected value and variance of r_{xy}^2 can be approximated by representing the function given in Equation 7 in terms of a partial Taylor series

expanded about $E(F'')$. This technique has also been used by Bobko and Rieck (1980) in the special case where $\beta_2 = 0$. The expected value and variance of r_{xy}^2 is approximated as follows:

$$E(r_{xy}^2) = g[E(F'')] + g''[E(F'')][Var(F'')/2] \tag{11}$$

$$Var(r_{xy}^2) = \{g'[E(F'')]\}^2 Var(F'') \tag{12}$$

where g is the function $F''/(N-2+F'')$,

$g[E(F'')]$ is the function g evaluated at the value $E(F'')$, and

$g'[E(F'')]$ and $g''[E(F'')]$ are the first and second derivatives of g evaluated at $E(F'')$.

To evaluate the expected value and variance of r_{xy}^2 as given by Equations 11 and 12, expressions are needed for the expected value and variance of the doubly noncentral F variable [$E(F'')$, $Var(F'')$]. Due to the complexity of a doubly noncentral F distribution, these parameters were approximated in terms of the moments of a central F distribution (Johnson & Kotz, 1970). The approximation consists of representing the doubly noncentral F with degrees of freedom 1 and $N-2$ and noncentrality parameters λ_1 and λ_2, as follows:

$$F'' = [F(d_1,d_2)](T) \quad , \tag{13}$$

where

$$T = (1+\lambda_1)/[1+\lambda_2/(N-2)] \quad , \tag{14}$$

$$d_1 = (1+2\lambda_1)^2/(1+4\lambda_1) \quad , \tag{15}$$

$$d_2 = (N-2+2\lambda_2)^2/(N-2+4\lambda_2) \quad , \tag{16}$$

and $F(d_1,d_2)$ = central F with degrees of freedom d_1 and d_2. The approximation given in Equation 13 has been shown to be reasonably accurate (Das Gupta, 1968; Patnaik, 1949).

Using the representation in Equation 13, the expected value and variance of the noncentral F'' variable can be expressed as

$$E(F'') = (T)E[F(d_1,d_2)] \tag{17}$$

$$Var(F'') = (T^2)Var[F(d_1,d_2)] \quad , \tag{18}$$

where $E[F(d_1,d_2)]$ and $Var[F(d_1,d_2)]$ are the mean and variance of an ordinary F distribution with degrees of freedom d_1,d_2 (Searle, 1971).

The final step is to substitute the results given in Equations 13, 17, and 18 into Equations 11 and 12 to obtain expressions for $E(r_{xy}^2)$ and $Var(r_{xy}^2)$. The same procedure (with some simple modifications) is used to obtain expressions for the expected

value and variance of the squared correlation in the selected group (r_{xys}^2) and the squared corrected correlation (r_{xyc}^2). The necessary modifications are as follows: (1) In deriving expressions for the expected value and variance of r_{xys}^2, all procedures are defined with respect to the selected group rather than the total group; (2) In deriving expressions for the expected value and variance of r_{xyc}^2, in addition to the modification given above, the function g described in Equations 9 and 10 is given as

$$g = F''/[(N-2)(s_{xs}^2/s_x^2)+F''] \tag{19}$$

Given expressions for the expected value and variance of r_{xy}^2, r_{xys}^2, and r_{xyc}^2, Equations 5 and 6 can be used to obtain expressions for the expected mean square error of the squared selected group (r_{xys}^2) and corrected (r_{xyc}^2) correlations as estimates of the population linear relationship.

Conditions and Procedures

The accuracy of r_{xys}^2 and r_{xyc}^2 was assessed by evaluating the EMSE of each estimator as a function of the form of the xy relationship (linear or nonlinear), the form of the marginal distribution of x scores, the strength of the xy relationship, sample size, and the proportion of the sample that is selected and measured on the y variable.

The form of the population relationship was defined in terms of the population regression equation, $\beta_0 + \beta_1 x + \beta_2 x^2$. One linear form and two nonlinear forms were investigated. For x scores in the range $[-3,3]$, the following regression forms were examined:

(a) $\beta_0 = 3$; $\beta_1 = 1$;
    $\beta_2 = 0$ (linear)
(b) $\beta_0 = 1.5$; $\beta_1 = 1$;
    $\beta_2 = .166$ (nonlinear, convex)
(c) $\beta_0 = 4.5$; $\beta_1 = 1$;
    $\beta_2 = -.166$ (nonlinear, concave)

In the convex case the regression equation flattens out as x decreases, whereas in the concave case the regression line flattens as x increases. Case b could occur if some floor effect was operating on the y variable. On the other hand, a ceiling effect could produce the concave regression line. It should

be noted that all three regression equations share the same functional values at x values of -3 and 3. However, at intermediate x values, the convex function lies below the linear function, and the concave function lies above. It should also be noted that the choice of $[-3,3]$ as the range for the x variable assures that the regression functions are monotonic (i.e., the slopes of the regression functions are never negative).

Four different marginal x distributions were employed: normal, uniform, skewed to the right, and skewed to the left. In the normal case the x scores were generated from a normal distribution with zero mean and unit variance. The uniform distribution of x scores was obtained by sampling from a uniform distribution in the interval $[0,1]$ and transforming the scores as $6x - 3$. The skewed distributions were defined by beta distributions with parameters $(5,10)$ for the right skewed case, and parameters $(10,5)$ for the left skewed case. As was done in the uniform case, the transformation $6x - 3$ was employed to assure that the x scores fell in the interval $[-3,3]$.

The strength of the xy relationship was defined in terms of the expected value of the squared correlation in the total group, $E(r_{xy}^2)$. The following values were chosen: .10, .30, .50. Total sample size was set at values of $N = 50$, 100, and 250. The proportion of cases selected (those scoring highest on x) was varied as follows: PSEL = .25, .50, .75.

By considering all possible combinations of the factors described above, a total of 324 conditions were investigated. In each case, the following procedures were employed:

1. Given a specification of the form of the x distribution and the sample size, a sample of N x scores was generated from the specified distribution using a pseudorandom number generator;
2. Given a specification for the form of the population regression equation (linear, convex, concave) and the strength of the xy relationship $[E(r_{xy}^2)]$, a value for the variance of y (σ_y^2) was chosen that yielded the desired $E(r_{xy}^2)$ value;
3. A specified proportion of the sample scoring highest on x was considered as selected;

4. Given the first three specifications, the EMSEs for the corrected and selected squared correlations were computed.

Results

As previously noted, the full study was a 5-factor factorial consisting of 324 conditions. For each of these conditions, two different measures were used to compared the accuracy of the correction formula (r_{xyc}^2) to that of the squared unselected group correlation (r_{xys}^2). PROP was defined as the proportion of times that r_{xyc}^2 yielded a smaller EMSE value than r_{xys}^2. DIFF was defined as the difference in the EMSE of the two estimates, computed as EMSE(r_{xys}^2) − EMSE(r_{xyc}^2). A positive value for DIFF means that the corrected value has a smaller EMSE value than the uncorrected value. The results are presented in Table 1. The marginal row and column means in Table 1 allow assessment of the main effects of each of the five factors. The individual cell means allow assessment of the interaction of the form of the regression with each of the other four factors. Both the cell means and the marginal row and column means were computed by averaging over the appropriate cells of the full 324-cell factorial.

The five main effects and four 2-factor interactions which are reported in Table 1 account for a substantial proportion of the between-cell variability in the full 5-factor factorial. For example, using PROP as the dependent variable, the effects presented in Table 1 account for 71% of the between-cell variability. The five main effects account for 56% of the variability, while the four 2-factor interactions account for 15%. Although higher-order interactions which involve the form of the regression function could also be investigated, these 3-, 4-, and 5-factor interactions would account for only an additional 8% of the between-cell variability. Thus, the decision was made to summarize the results in the form given in Table 1.

The overall effect of the form of the regression on the relative accuracy of the correction formula can be clearly seen by examining the column marginals. Two different results can be observed:

1. The relative accuracy of the correction formula is similar in the linear and concave cases. The correction formula yields a smaller EMSE value than the squared selected group correlation in 52% of the cases when the regression is linear, and in 57% of the cases when the regression is of the nonlinear concave form.

2. The relative accuracy of the correction formula greatly deteriorates for the nonlinear convex condition. In this case, the correction formula is the superior estimate in only 15% of the cases. Even poorer results are seen for a convex regression form when the x distribution is skewed to the right, or when the strength of the linear xy relationship is at its lowest value of E(r_{xy}^2) = .10.

Under both of these conditions, there are no instances where the correction formula has a smaller EMSE value than the selected group estimate. Further, when the regression is convex in form, there is not a single instance where the correction formula is the desirable estimator in 50% or more of the cases.

It is also of interest to note that even when the regression forms are linear or concave, the percentage of cases where the correction formula is superior to the squared selected group correlation is surprisingly small. For example, for the linear regression condition, it is seen that when the population xy relationship is relatively low (i.e., E(r_{xy}^2) = .10), the correction formula never has a smaller EMSE value than the selected group estimate. In the concave case, the correction formula is superior only 19% of the time. It is not until the population xy relationship is quite high, E(r_{xy}^2) = .50, that the correction formula clearly dominates. Similar results can be seen when the sample size and proportion selected factors are considered. For the linear and concave regression forms, the correction formula is clearly superior to the selected sample estimator only for large sample sizes (N = 250) or when 75% of the cases are selected.

Conclusions

It was found that although the relative accuracy of the correction formula is similar under both a linear and nonlinear concave regression form (i.e., a regression line that flattens as x increases), the

Table 1
Accuracy of the Uncorrected and Corrected Squared Correlations
as a Function of the Form of the Relationship, the Marginal
Distribution of x, the Strength of the Relationship, the Size
of the Total Group, and the Proportion of Cases Selected

| Variable | Accuracy | Form of the Relationship | | | Mean |
		Linear	Convex	Concave	
Distribution of x					
Normal	Prop	.52	.07	.59	.40
	Diff	.003	-.019	.002	-.005
Uniform	Prop	.48	.11	.48	.36
	Diff	.005	-.022	-.008	-.008
R-Skew	Prop	.52	.00	.63	.38
	Diff	.001	-.021	.007	-.004
L-Skew	Prop	.56	.41	.59	.52
	Diff	.008	-.004	.001	.001
Mean	Prop	.52	.15	.57	
	Diff	.004	-.017	.001	
Strength of the Relationship					
.10	Prop	.00	.00	.19	.06
	Diff	-.035	-.041	-.030	-.035
.30	Prop	.56	.06	.64	.42
	Diff	-.005	-.019	-.007	-.010
.50	Prop	1.00	.39	.89	.76
	Diff	.052	.001	.039	.034
Mean	Prop	.52	.15	.57	
	Diff	.004	-.017	.001	
Size of the Total Group					
50	Prop	.36	.14	.36	.29
	Diff	-.010	-.019	-.025	-.018
100	Prop	.56	.14	.53	.41
	Diff	.003	-.019	.002	-.005
250	Prop	.64	.17	.83	.55
	Diff	.020	-.012	.025	.011
Mean	Prop	.52	.15	.57	
	Diff	.004	-.017	.001	
Proportion of Cases Selected					
.25	Prop	.42	.22	.31	.31
	Diff	-.007	-.026	-.032	-.022
.50	Prop	.56	.11	.64	.44
	Diff	.012	-.018	.017	.004
.75	Prop	.58	.11	.78	.49
	Diff	.008	-.006	.017	.006
Mean	Prop	.52	.15	.57	
	Diff	.004	-.017	.001	

performance of the correction formula is poor relative to the selected group estimate when the regression function is of the nonlinear convex form (i.e., a regression line which flattens as x decreases). Under the convex condition, the uncorrected selected group squared correlation yielded a smaller EMSE value in 85% of the cases investigated. This result can best be understood by considering the bias of each estimator.

When the regression line is convex, and the persons scoring highest on x are selected, the slope of the regression line in the selected group will

overestimate the actual linear slope coefficient. Consequently, the correction formula will tend to strongly overestimate the linear xy relationship. The selected group estimate, although negatively biased in this case due to the restriction in range of the x variable, will have a smaller absolute bias value than the correction formula. In support of this argument, it was found that the average bias values of the corrected and uncorrected estimates for a convex regression form were .11 and $-.07$, respectively.

Similarly, the improved performance of the correction formula under a concave regression form can also be understood in terms of the bias of the estimates. Given a concave regression form, the selected group estimate of the slope coefficient will be underestimated due to the flattening of the regression slope in the selected group. In this case, both the correction formula and the selected group estimate will tend to be negatively biased, with the former having the smaller absolute bias. It was found that for the concave regression function the average bias of the corrected and selected group estimates was $-.10$ and $-.19$, respectively.

It should also be noted that even when the regression of y on x is linear or concave in form, the correction formula is not necessarily the desired estimator. It has been observed that unless the linear xy relationship is relatively high ($E(r_{xy}^2) \geq .30$) or sample size or the proportion selected are relatively high ($N \geq 100$, PSEL $\geq .50$), the correction formula was actually inferior to the uncorrected estimate more than 50% of the time.

Finally, the degree of nonlinearity considered in the present study could at best be described as moderate. First, the nonlinear regression functions were restricted to be monotonic in form. Thus, for example, in the convex case U-shaped functions were not investigated. Secondly, the coefficient for the quadratic term in the regression equation was not large, being set to a value of $\beta_2 = \pm.166$. It is reasonable to assume that even less favorable results for the correction formula would have been obtained if stronger degrees of nonlinearity had been considered. The results of the present research suggest that in many instances, the best strategy may be to simply report the uncorrected correlation.

References

Bobko, P., & Rieck, A. (1980). Large sample estimators for standard errors of functions of correlation coefficients. *Applied Psychological Measurement, 4,* 385–398.

Das Gupta, P. (1968). Two approximations for the distribution of double noncentral beta. *Sankhya Series B, 30,* 83–88.

Greener, J. M., & Osburn, H. G. (1979). An empirical study of the accuracy of corrections for restriction in range due to explicit selection. *Applied Psychological Measurement, 3,* 31–41.

Greener, J. M., & Osburn, H. G. (1980). Accuracy of corrections for restriction in range due to explicit selection in heteroscedastic and nonlinear distributions. *Educational and Psychological Measurement, 40,* 337-346.

Gross, A. L. (1982). Relaxing the assumptions underlying corrections for restriction of range. *Educational and Psychological Measurement, 42,* 795–801.

Gross, A. L., & Fleischman, L. (1983). Restriction of range corrections when both distribution and selection assumptions are violated. *Applied Psychological Measurement, 7,* 227–237.

Johnson, N. L., & Kotz, S. (1970). *Continuous univariate distributions* (Vol. 2). Boston: Houghton Mifflin.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Patnaik, C. B. (1949). The noncentral χ^2 and F distributions and their applications. *Biometrika, 61,* 202–232.

Searle, S. R. (1971). *Linear models.* New York: Wiley.

Author's Address

Send requests for reprints or further information to Alan L. Gross, Department of Educational Psychology, Graduate Center, City University of New York, 33 West 42nd Street, New York NY 10036, U.S.A.