Order Number 8821121

A technique for the treatment of missing data in a nonlinear regression model

Shulman, Vivian Gross, Ph.D.

City University of New York, 1988

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy. Problems encountered with this document have been identified here with a check mark ✓.

1. Glossy photographs or pages _____
2. Colored illustrations, paper or print _____
3. Photographs with dark background _____
4. Illustrations are poor copy _____
5. Pages with black marks, not original copy ✓
6. Print shows through as there is text on both sides of page _____
7. Indistinct, broken or small print on several pages ✓
8. Print exceeds margin requirements _____
9. Tightly bound copy with print lost in spine _____
10. Computer printout pages with indistinct print _____
11. Page(s) _____ lacking when material received, and not available from school or author.
12. Page(s) _____ seem to be missing in numbering only as text follows.
13. Two pages numbered _____. Text follows.
14. Curling and wrinkled pages _____
15. Dissertation contains pages with print at a slant, filmed as received ✓
16. Other _____

_____

_____

A TECHNIQUE FOR THE TREATMENT OF MISSING DATA IN A NONLINEAR REGRESSION MODEL

by

VIVIAN G. SHULMAN

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1988

This manuscript has been read and accepted for the Graduate Faculty in Educational Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

4/25/88 — Date

Alan J. Gross — Chair of Examining Committee

4/25/88 — Date

Carol Kehr Tittle — Executive Officer

Dr. Alan L. Gross, Chairman

Dr. David M. Rindskopf

Dr. Barry J. Zimmerman
Supervisory Committee

The City University of New York

Abstract

A TECHNIQUE FOR THE TREATMENT OF MISSING DATA IN A NONLINEAR REGRESSION MODEL

by

VIVIAN SHULMAN

Adviser: Prof. Alan L. Gross

A problem occurs in educational practice when an organization wishes to validate a test, x, as a predictor of a criterion variable, y, and the data sets are incomplete. Often, due to the selection of subjects on the basis of x, criterion (y) scores are available for some subjects, but missing for others. A statistical problem of interest is to estimate the missing y scores in an attempt to infer the relationship between x and y in the total group. Regression techniques for handling this problem assume a linear regression model. The problem is exacerbated in the fairly frequent case when the regression of y on x is nonlinear in form.

The primary goal of this research was to analytically investigate the effectiveness of three regression techniques for estimating missing y scores, when the underlying model was nonlinear, and

specifically quadratic in form. Method 1 simply utilized the x scores in the selected group to predict missing y scores. Method 2 utilized an auxiliary variable, z, in conjunction with x to predict missing y values. And method 3, a special case of method 2, utilized x and x^2 to predict missing y values. Expressions to compare the expected mean squared error of each of the three regression procedures were analytically derived. These expressions were compared in terms of sample size, the proportion of cases selected, the distribution of x scores, the relationship of y to x, and the relationship of z to x. The findings of the present study indicate that first, as expected, in the case where the underlying xy relationship is linear, the simplest regression method (i.e., utilizing the selected x cases alone) performs best in predicting the missing y cases. Second, in a situation where the relationship between x and y is assumed to be non-linear, the utilization of an additional variable in conjunction with x is the method of choice in predicting missing y cases. Finally, in a situation where the range of x values is severely restricted, the performance of all three procedures is unreliable, and the performance of procedure 3 is especially poor. Recommendations for researchers, and potential areas for future research are discussed.

ACKNOWLEDGEMENTS

It is my pleasure to acknowledge Dr. Alan L. Gross, who served as my dissertation advisor. Dr. Gross' astute thinking, patience, and guidance made this investigation possible.

I would also like to thank the other members of my dissertation committee, Dr. David M. Rindskopf and Dr. Barry J. Zimmerman, for their invaluable suggestions regarding the study.

Finally, I would like to thank Steven S. Gross for his help in writing the computer program and for his patience and support.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

A frequently occurring problem in educational practice arises when one wants to validate a test, x, as a predictor of some criterion variable, y, and the x-y data sets are incomplete in that x or y (or both) is missing for some subjects. A special case of this missing data problem, known as restriction of range, commonly occurs as a result of selecting subjects on the basis of x. Thus, y scores are known for the selected cases, but are unknown for the unselected cases. A statistical problem of interest is to estimate the missing y scores in an attempt to infer the total group x-y relationship.

When a regression function is computed on a subset of data (for example: x and y scores for those selected on x), it may be difficult to generalize the obtained equation to the full set of scores. Consider the following example which has received much attention in the literature. Suppose a college selects a certain proportion of students scoring highest on a test, x, for

admittance. The institution is then interested in validating their admissions procedure by observing the relationship between X and scores on a criterion variable, y. Let x represent a college entrance examination such as the Scholastic Aptitude Test, and let y represent first year grade point average (G.P.A.). While x scores are observed for all applicants, y is observed for selected cases only (those cases scoring in the highest p percent), and missing for the unselected cases.

Typically, regression techniques for handling this type of missing data problem are based on the assumption of a linear regression model. The techniques rely on extrapolation, where the linear regression of y on x is obtained for a selected group, and extrapolated back for the full data set. This type of extrapolation can be problematic when the regression of y on x is nonlinear, a not infrequent case. In our example, suppose the regression of GPA scores on SAT scores is concave in form, with GPA scores flattening out as SAT scores increase. In this case, a linear regression analysis performed with the selected data will yield a distorted view of the true regression function. Further, this may yield inaccurate estimates of the missing Y scores of the unselected cases.

We now contemplate the problem in more detail. Consider a situation where there is a quadratic relationship between x and y in the population. The population regression function can be expressed by the following equation:

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + e \qquad (1)$$

Given this nonlinear underlying model described by equation (1), suppose one cannot observe both x and y for all N subjects; due to selection on x, there are missing data on y. We are interested in estimating the missing y scores for the unselected subjects. An approach that is often used in dealing with this missing data problem is to apply a linear least squares analysis to the subset of complete cases (i.e., the cases for which both x and y are observed) (Yates, 1933), and use this equation to estimate missing y scores. This approach will here and after be referred to as "procedure 1." The linear regression of y on x that results from this procedure can be expressed as follows:

$$\tilde{Y}_S = \tilde{B}_{0S} + \tilde{B}_{1S}X \qquad (2)$$

where $\underline{B}_S = \begin{bmatrix} \tilde{B}_{0S} \\ \tilde{B}_{1S} \end{bmatrix} = (X_S' X_S)^{-1} X_S' \underline{Y}_S$

and X_S is an N_S by 2 matrix, whose first column is the unit vector and whose second column contains the x scores for the selected cases. We can then utilize equation (2) by applying it to the x scores of the unselected cases in order to obtain the missing y scores.

An illustration of the problem that may result when a linear regression analysis is performed on the subset of complete cases is demonstrated in Figure 1.

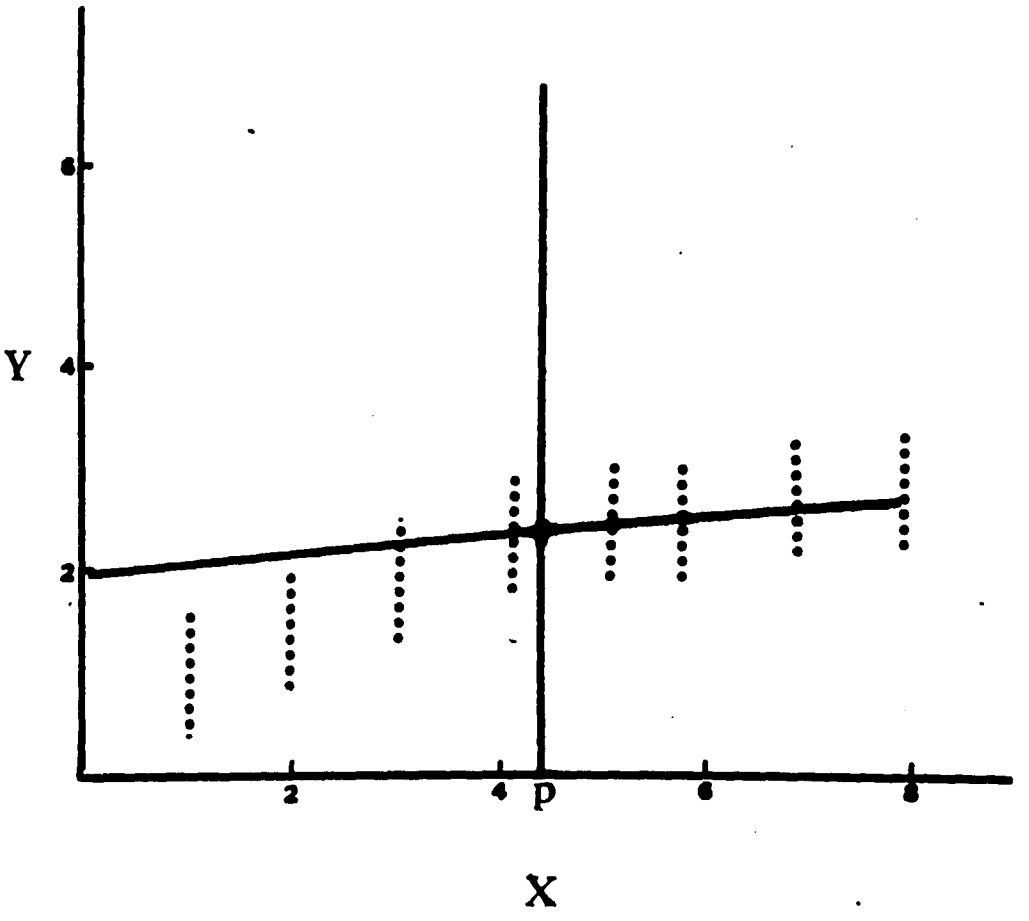


Figure 1. A hypothetical xy data set[a].

[a]p = at a given x value, a proportion, p, is selected.

Let us consider the hypothetical data set depicted in Figure 1. Suppose we only had scores to the right of the perpendicular line, P. An attempt to extrapolate back linearly from these scores to estimate the scores on the left side of the line will lead to an overestimation of points.

Given the underlying nonlinear model expressed by equation (1), and missing data on y due to selection on x, we consider an alternate approach to the missing data problem, i.e., to the estimation of the missing y scores. This method involves the introduction of an additional variable, z, where complete data on z is available. We assume that given x and x^2, z is independent of y, but given x, y and z may be dependent. The y variable is regressed on x and z, using the x, y, and z data for the selected cases. The resulting least squares weights (W_0, W_1, W_2,)are applied to x and z in the unselected group to predict the missing y scores. The predicted values are denoted as $\hat{y}$, and are computed from the following equation:

$$Y = W_0 + W_1X + W_2Z \qquad (3)$$

This procedure will here and after be referred to as Procedure 2. A potential advantage to procedure 2 is that the utilization of the variable z in conjunction with x will promote a good approximation of the correct model in the prediction of missing y cases.

Given the underlying nonlinear model expressed in equation (1), and missing data on y due to selection on x, a third method for estimating missing y cases is presented. This procedure is a special case of Procedure 2, where the variable z is set equal to x^2. Thus, with this method, the correct model is utilized, where y is regressed on x and x^2. However, it should be noted that we are utilizing the variables x and x^2 for the range of selected subjects only.

This procedure (referred to as Procedure 3), proceeds as follows: The y variable is regressed on x and x^2 in the selected group. The resulting least squares weights ($\tilde{\alpha}_0$, $\tilde{\alpha}_1$, $\tilde{\alpha}_2$) are applied to x and x^2 in the unselected group to predict the missing y scores. The predicted values are denoted as $\hat{Y}$, and are computed from the following equation:

$$Y = \tilde{\alpha}_0 + \tilde{\alpha}_1 X + \tilde{\alpha}_2 X^2 \qquad (4)$$

where $\underline{\tilde{\alpha}}$ is computed in the selected group.

One can evaluate the accuracy of each of the three procedures by analyzing how well equations 2, 3, and 4 estimate the missing Y cases. More specifically, we can consider the expected mean squared error (EMSE) of prediction of each of the three procedures. A prototype expression for these indices can be illustrated as follows:

$$EMSE = \frac{E\sum_{i=1}^{N_u}(Y_i - \hat{Y}_i)^2}{N_u} \qquad (5)$$

where N_u = the number of cases in the unselected group.

Y_i = the actual y score for the i^{th} subject in the unselected group.

$\hat{Y}_i$ = the estimated y scores obtained by one of the three procedures for the i^{th} subject in the unselected group.

In the present study, expressions for the accuracy of the EMSE of the three procedures are analytically derived. We will compare these expressions in terms of the following parameters: a) sample size, b) proportion of cases selected, c) the distribution of x scores, d) the relationship of y to x, and e) the relationship of z to x.

Results from this study can be valuable to educational practitioners. Very often when investigating relationships among sets of scores, one may encounter the phenomenon known as a "ceiling effect," where a regression line will level off at the higher end of a curve. There are several explanations why this happens. Often the maximum obtainable score on a test is a good deal lower than 100% correct. Sometimes a test is too easy (or too difficult) leaving no room for improvement. Given such a situation, coupled with missing data on a variable, a more accurate regression might be obtained from generalizing results from this study.

Chapter 11

REVIEW OF THE LITERATURE

Much of the research on missing data has focused upon the problem of restriction of range, where the main emphasis is on correction formulae estimates and their properties. For a review of the research in this area one can consult the following studies: Pearson (1903), Lawley (1943), Cohen (1955), Novick and Thayer (1968), Kagan (1977), Gross (1981,1982), Linn (1983), Greener and Osburn (1980), Gross and Fleischman (1983), Roe (1979), Olson and Becker (1983).

The following review of the literature will focus on the broader problem of missing data in multivariate data analysis, and specifically regression analysis. The different procedures for dealing with missing data when performing multiple regression analyses are discussed together with the distributional assumptions underlying these methods. Properties of these procedures are compared through a review of simulation studies. The procedures discussed in the literature are intended for use on data matrices whose entries are missing by virtue of some process that is unrelated to any of the relationships between the variables in the matrix. Thus, there is an underlying assumption that the

process of selection on one variable is independent of the other variables.

There are two main approaches for estimating missing data. The first is a statistical procedure which begins by assuming that the observed data is a sample drawn from a multivariate distribution of a known form (usually multivariate normal), and having unknown parameters. The population parameters are estimated using the available data and an estimation procedure (such as maximum likelihood). The estimated parameters are then used to estimate particular missing observations, by utilizing the conditional distribution of the variable whose observations are missing, given the variables whose observations are not missing. Orchard and Woodbury (1972) estimate the parameters of a multivariate normal distribution utilizing the method of maximum likelihood. Beale and Little (1975) derive Orchard and Woodbury's "missing information principle," and demonstrate how the principle leads to a simple iterative algorithm for finding estimators that are maximum likelihood when the population is multivariate normal.

Gleason and Staelin (1975) indicate that the major advantage of utilizing a statistical method in dealing with missing data is that it yields "an unambiguous model of the estimation process whose characteristics

can be evaluated analytically." However, this method is based upon stringent assumptions about the distributional form of the population. Very often the assumption of a multivariate normal distribution is inappropriate. For example, questionnaires often consist of items having only several responses and result in data which are clearly not normal.

To avoid the distribution assumptions of the statistical approach, the problem of estimating missing data is often approached by a second method, which does not model the generation of data. Instead, the information in variables having no missing values is utilized to construct reasonable estimates for missing entries. Results from this approach are evaluated not by studying the properties of the model, but by measuring how well the various methods can reconstruct unknown values from available data using actual examples. Typically, such research utilizes a multiple linear regression model, where data is missing at random from one or more variables, and the population is assumed to have a multivariate normal distribution.

Yates (1933) proposed the application of the usual least squares method to all available observations; i.e., the subset of complete cases. However, unless the amount of data is very small, the results can be very unsatisfactory. This is particularly true if the value

of many variables are known for an incomplete observation, and these variables are important for the study. Beale and Little (1975) demonstrate that the leastt squares method is inferior to all other alternatives, where their criterion for estimating the effectiveness of other estimators is the "residual sum of squares of deviations of the observed and fitted values of the dependent variables when the deleted values are restored" (p. 139). (The Beale and Little report will be discussed in a later section, where the criterion described above for judging the effectiveness of estimators is elaborated upon).

Another approach to this missing data problem is to substitute suitable guessed values for the unknown quantities. Wilks (1932) proposed using the mean of all nonmissing values for any variable to estimate the missing values. However, with data that is highly correlated, this can give very poor results (Beale and Little, 1975; Haitovsky, 1966). Gleason and Staelin (1975), Affifi and Elashoff (1966), and Timm (1970) investigated the effectiveness of the Wilks method along with other missing data methods. They conclude that if the average intercorrelation among variables is in excess of .2, then the redundancy in the variables can be effectively used to estimate values of missing entries. Only when the intercorrelation of variables is

below .2 will the best information about a missing entry come only from the variable itself (i.e., the mean).

Another method, proposed by Dear (1959), to retrieve missing data uses the property of principal component analysis, namely, that the original data are obtainable from factor scores and factor loadings. Dear's method decomposes the data matrix into its known and unknown parts and uses the first principal component and its associated loadings derived from the known data to estimate the unknown data.

Finally, a method proposed by Buck (1960) to replace missing data uses a regression equation to compute estimates for missing entries. Buck begins by using the complete observations to estimate the means of all the variables and the covariance matrix. These values are then used to estimate any missing quantities as linear functions of the variables that are known for the observations. Thus, the independent variables are the nonmissing entries for the individual who has a missing entry, and the regression coefficients are calculated using all subjects in the submatrix which has no missing entries.

The missing data techniques of Dear, Wilks, and Buck are compared in a study by Timm (1970), who used as a criterion the ability of each method to predict the correlation or covariance matrix. The Dear and Buck

methods were generally superior to the Wilks method for the data matrices that were investigated. Timm's results, however, provide no information on the problem of estimating the missing entries themselves.

Beale and Little (1975) compared the properties of several missing data procedures, including the method of ordinary least squares on complete cases only, and Buck's method, utilizing a simulation study. A third procedure considered by Beale and Little is an iterated form of Buck's method. Buck (1960) used only the complete observations to estimate the means of all the variables and the covariance matrix. The iterated form of Buck's procedure takes trial values for the means (X_j) of the variables and the covariance matrix, and uses them to estimate the missing quantities using linear regression. The process is repeated, each time resetting the value for the means and estimated covariance matrix, until there are no further changes. Beale and Little note that Orchard and Woodbury derive an algorithm that produces the same maximum likelihood estimates as the iterated version of Buck's method, when the population is multivariate normal. (The only difference is that the adjusted sum of squares and products matrix is divided by (N-1) instead of (N), to derive the estimated covariance matrix). An advantage of the iterated Buck procedure over the Orchard and

Woodbury maximum likelihood method is that it does not assume that the underlying population is multivariate normal.

In their study, Beale and Little generated data from a multivariate normal population with one variable identified as the dependent, and between two and four independent variables. Five, ten, twenty and forty percent of the observed values for each variable were randomly deleted. The criterion utilized for judging the effectiveness of each procedure was the "residual sum of squares of deviations of the observed and fitted values of the dependent variable when the deleted values were restored" (p. 139). This can be written in symbols as

$$S = \sum_{i=1}^{N} \left\{ Y_i - b_0 - b_j X_{ij} \right\}^2 \qquad (6)$$

where b_0 and b_j are the constant term and slope coefficient estimated from incomplete data by one of the methods; and X_{ij} and Y_i are the true values of all variables having no deletions.

Beale and Little conclude from their analyses that Buck's method and iterated Buck perform consistently better (produce lower values of S) than ordinary least squares on complete cases only. While the iterated Buck

procedure requires more computing than the ordinary Buck method, the former always performed better than the latter, except for three cases having only 5% deletions.

While the properties of the missing data procedures described above have been compared in simulation studies, analytic comparisons have generally not been performed. This is due to the complexity of the problem that arises when missing values are scattered over several explanatory variables. Donner (1982) attempts to overcome this difficulty by focusing on the simple case of two explanatory variables (x_1, x_2) with missing values on x_2 only, where the regression of y on x_1 is assumed linear. Using the model $Y = b_0 + b_1X_{1i} + b_2X_{2i} + e$; i=1,2...N, Donner derives algebraic expressions for the bias and variance associated with various methods of estimating the parameters b_1 and b_2. The errors (e_i) are assumed to be independently distributed with zero means and unit variances over all N cases. Included among the procedures that Donner investigates are: a) the complete case method, where the usual least squares method is applied to the N cases for which x_1 and x_2 are observed; b) the mean substitution method (Wilks, 1932), where a missing value on x_2 is replaced by the mean of x_2 over the complete cases; c) the method of linear prediction (Buck, 1960), where a missing x_{2i} is replaced by its "predicted value from the simple

linear regression of x_2 on x_i over the complete cases" (p. 379).

As a result of his analysis, Donner concludes that the mean substitution procedure is relatively effective in estimating the coefficients for incompletely observed variables when the intercorrelations among the variables are weak, and the proportion of missing cases fairly high. However, this method tends to be biased if the correlations among the variables are high, and the bias persists even with large samples having random patterns of missing data.

When comparing the linear prediction method with the complete case method, Donner found that the former method is more accurate than the latter in estimating b_1, while the two procedures were equivalent for estimating b_2.

In summary, the studies that have been discussed assume that the relationship among variables is linear in form. The following study differs from the previous research in that we investigated the problem of missing data when the relationship among the variables was nonlinear. A simple but realistic case of missing data on y due to selection on x was investigated. Three procedures were compared analytically. Procedure 1 regressed y on x in the selected group, and applied the obtained regression coefficients to the x scores of the

unselected group in order to obtain missing y scores. Procedure 2 regressed y on x and z in the selected group, and applied the resulting least squares weights to x and z in the unselected group to predict missing y scores. Procedure 3 regressed y on x and x^2 in the selected group, and applied the resulting least squares weights to x and x^2 in the unselected group to predict missing y scores.

Chapter 111

METHOD

The investigation of the three procedures for dealing with missing data in a nonlinear regression model was conducted from an analytic vantage point. Three indices were derived to evaluate the three procedures. Let us briefly review procedures 1,2, and 3. In procedure 1, y is regressed on x in the selected group. Obtained regression weights are applied to x scores in the unselected group, and subsequently the missing y scores in the unselected group are estimated. In procedure 2, y is regressed on x and z in the selected group. Resulting regression weights are applied to x and z scores in the unselected group to obtain missing y scores. In procedure 3, y is regressed on x and x^2 in the selected group. Resulting least squares weights are applied to x and x^2 in the unselected group to predict missing y scores.

Indices 1,2, and 3 (which correspond to procedures 1,2,and 3 respectively), gauge how well the three procedures can reproduce the actual y scores for the unselected cases. The three indices represent the

expected mean squared error (EMSE) of prediction for each of the three procedures.

Analytic evaluation of procedures 1,2, and 3 occurred in two steps. Step one involved the derivation of mathematical expressions for indices 1,2, and 3. Step 2 involved evaluating the performance of the three procedures by varying the three indices with respect to various conditions.

Let us now consider the mathematical derivations for indices 1,2, and 3. A prototype expression for the expected value of the EMSE for the three indices is illustrated as follows:

$$\text{EMSE} = \frac{E(\underline{Y}_u - \hat{\underline{Y}}_u')(\underline{Y}_u - \hat{\underline{Y}}_u')}{N_u} \qquad (7)$$

where $\underline{Y}_u$ = the actual vector of y scores for the unselected cases.

$\hat{\underline{Y}}_u$ = the predicted vector of y scores for the unselected cases, utilizing one of the three procedures described above for estimating the missing scores.

N_u = the number of cases in the unselected group.

Let us consider the expression of the estimated y scores in the unselected group (i.e., $\hat{\underline{Y}}_u$). For

procedure 1, the estimated y scores in the unselected group can be expressed as follows:

$$\hat{Y}_{u1} = X_{u1}\hat{B}_{s1}$$

where $\hat{Y}_{u1}$ = the estimated y scores in the unselected group according to procedure 1.

X_{u1} = the N_u by 2 data matrix consisting of the X_0 and X scores of the unselected cases.

$$X_{u1} = \begin{matrix} & X_0 & X \\ \begin{matrix}1\\2\\\vdots\\N_u\end{matrix} & \left[\begin{matrix} & \\ & \\ & \end{matrix}\right. & \left.\begin{matrix} \\ \\ \end{matrix}\right] \end{matrix}$$

$\hat{\underline{B}}_{s1}$ = A vector of regression weights, where y in the selected group is regressed on X_0 and X in the selected group.

Now $\hat{\underline{B}}_{s1}$ can be expressed as follows:

$$\hat{\underline{B}}_{s1} = (X'_{s1} X_{s1})^{-1} X'_{s1} \underline{Y}_s$$

where X_{s1} = the N_s by 2 data matrix consisting of the X_0 and X scores for the selected cases.

$\underline{Y}_s$ = a vector of y scores for the selected cases.

Let us now consider the expression for the estimated y scores in the unselected group for procedure 2. This expression can be written as follows:

$$\hat{Y}_{u2} = X_{u2}\hat{B}_{s2}$$

where $\hat{Y}_{u2}$ = the estimated y scores in the unselected group, derived according to procedure 2.

X_{u2} = the N_u by 3 data matrix, consisting of the X_0, X and Z scores for the unselected cases.

$$= \begin{array}{c} \\ 1 \\ \vdots \\ N_u \end{array} \begin{array}{c} \begin{array}{ccc} X_0 & X & Z \end{array} \\ \left[\begin{array}{ccc} & & \\ & & \\ & & \end{array}\right] \end{array}$$

$\underline{B}_{s2}$ = a vector of regression weights predicting y in the selected group from X_0, X, and Z in the selected group.

$\underline{B}_{s2}$ can be expressed as follows:

$$\underline{B}_{s2} = (X_{s2}' X_{s2})^{-1} X_{s2}' Y_s$$

where X_{s2} = the N_s by 3 data matrix consisting of X_0, X and Z scores for the selected cases.

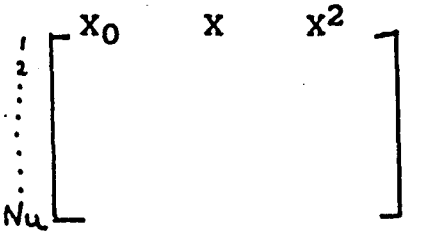
Y_s = the vector of y scores for the selected group.

Let us now consider the expression for the estimated y scores in the unselected group for procedure 3. This expression can be written as follows:

$$\hat{Y}_{u3} = X_{u3}\hat{\underline{B}}_{s3}$$

where $\hat{Y}_{u3}$ = the estimated y scores in the unselected group derived according to procedure 3.

X_{u3} = the N_u by 3 data matrix consisting of the X_0, X and X^2 scores for the unselected cases.

X_{u3} = 

$\hat{B}_{S3}$ = a vector of regression weights predicting y in the selected group from X_0, X, and X^2 in the selected group.

$\hat{B}_{S3}$ can be expressed as follows:

$$\hat{B}_{S3} = (X_{S3}' X_{S3})^{-1} X_{S3}' Y_S$$

where X_{S3} = the N_S by 3 data matrix consisting of the X_0, X and X^2 scores for the selected cases.

Y_S = the vector of y scores for the selected group.

In general, one can express a typical set of predicted y scores as follows:

$$\hat{Y} = GY_S$$

where $\hat{Y}$ = a vector of N_u predicted y scores.

Y_S = a vector of N_S predicted y scores.

G = an N_u by N_s matrix, whose elements are used in conjunction with $\underline{y}_s$ to obtain predicted y scores.

More specifically, G can be described as follows:

$$G = X_u (X_s' X_s)^{-1} X_s'$$

It follows that a typical index to evaluate the Expected Mean Squared Error (EMSE) of any of the three procedures is expressed as follows:

$$EMSE = \frac{E(\underline{y}_u - G\underline{y}_s)' (\underline{y}_u - G\underline{y}_s)}{N_u}$$

This expression can be written in the following form:

$$\frac{E(\underline{y}_u - G\underline{y}_s)' (\underline{y}_u - G\underline{y}_s)}{N_u} = \frac{E\ \underline{y}' A\ \underline{y}}{N_u}$$

where A is the following N by N matrix:

$$A = \begin{bmatrix} I & -G \\ -G & G\ G \end{bmatrix} \qquad (8)$$

where I = an N_u by N_u identity matrix.

$$G = X_u (X_s' X_s)^{-1} X_s'$$

and $\underline{y}$ = a vector containing the y scores of the selected and unselected cases;

$$= \begin{bmatrix} \underline{y}_u \\ \underline{y}_s \end{bmatrix}$$

If $\underline{y}$ is normally distributed, with expectation, $\underline{\mu}$, and a variance covariance matrix, Σ ;

then $\frac{E\underline{y}'A\,\underline{y}}{N_u}$ can be expressed as follows: (Searle, 1971, p. 55)

$$\frac{E(\underline{y}'A\,\underline{y})}{N_u} = \frac{tr\,(A\Sigma) + \underline{\mu}\,A\,\underline{\mu}}{N_u}$$

In the problem being considered:

$$\Sigma = \sigma^2 I$$

$$\underline{\mu} = X\,\underline{\alpha}$$

where $\sigma^2 I$ = the variance of y multiplied by an N by N identity matrix.

X = an N by 3 data matrix consisting of X_0, X and X^2.

$$X = \begin{bmatrix} X_0 & X & X^2 \\ & & \\ & & \end{bmatrix}$$

$\underline{\alpha}$ = a vector of regression weights having the following form:

$$\underline{\alpha} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix}$$

A typical index can therefore be written as follows:

$$\frac{E\,(\underline{y}_u - G\underline{y}_s)'(\underline{y}_y - G\,\underline{y}_s)}{N_u} = \frac{\sigma^2 tr(A) + \underline{\alpha}'X'A\,X\underline{\alpha}}{N_u}$$

where A is as in (8), and where G will be different for each of the three indices:

for procedure 1, $G = X_{u1}(X_{s1}'X_{s1})^{-1}X_{s1}'$

for procedure 2, $G = X_{u2}(X_{s2}'X_{s2})^{-1}X_{s2}'$

for procedure 3, $G = X_{u3}(X_{s3}'X_{s3})^{-1}X_{s}'$

Step two in the analytic evaluation of the three procedures involved the construction of a Fortran computer program. Subsequently, the performance of the three procedures was evaluated by varying the three indices with respect to the following conditions:

1) sample size: N = 25, 50, 100.

2) the amount of missing data: The proportion of selected data (PSEL) was varied as follows: .60, .75, .90. The data that were selected were always at the upper end of the distribution.

3) the form of the regression of y on x: Three forms were used: linear; non linear but monotonic; non linear and non-monotonic.

4) the form of the regression of z on x: Three forms were used: non linear but monotonic; non linear and non-monotonic.

5) <u>the distribution of x scores:</u> Three types of x score distributions were considered: a) uniform, b) skewed to the left, and c) skewed to the right.

We will now elaborate on conditions 3,4, and 5 by discussing the manner in which the x, y, and z data were generated, and the interrelationships among the variables.

The x variable was constructed to have a range from 1 to 10. Three different distributions of x score values were constructed, namely, uniform, skewed to the left, and skewed to the right. Table 1 demonstrates the percentage of cases distributed at each value of x, for each of the three different distributions.

Table 1

The Distribution of x Scores by Distribution and Sample Size

x Scores	x Distribution								
	Uniform			Left-skewed			Right-skewed		
	30	50	100	25	50	100	25	50	100
1	3	5	10	1	2	4	8	16	32
2	3	5	10	1	2	4	5	10	20
3	3	5	10	1	2	4	3	6	12
4	3	5	10	1	2	4	2	4	8
5	3	5	10	1	2	4	2	4	8
6	3	5	10	2	4	8	1	2	4
7	3	5	10	2	4	8	1	2	4
8	3	5	10	3	6	12	1	2	4
9	3	5	10	5	10	20	1	2	4
10	3	5	10	8	16	32	1	2	4

Let us examine the three types of x score distributions more closely, beginning with x scores that are uniformly distributed. Figure 2 illustrates a case where x scores are uniformly distributed between 1 and 10.

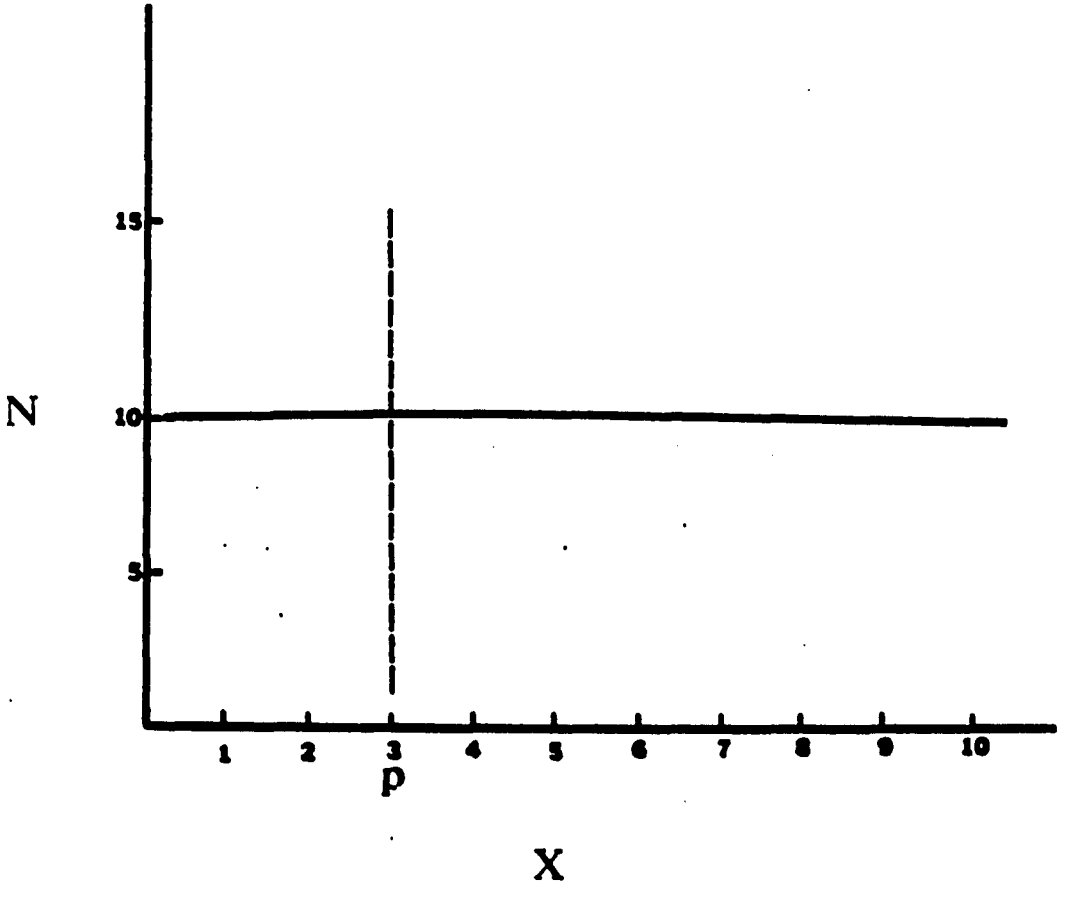


Figure 2. A uniform distribution of x scores[a].

[a]p = at a given x value, a proportion, p, is selected.

The solid horizontal line illustrates a distribution of x scores that is uniform. The dotted vertical line (p) represents the proportion of cases selected. If we consider this uniform distribution of x scores, and assume that the total number of cases (N) is 100, then 10 x scores would occur at each x value from 1 to 10. If the proportion of data selected (psel) is the upper 75% of cases, then the lower 25 cases will be deleted. Therefore, in the case where scores are uniformly distributed, one would be missing all x cases in the range of 1 and 2, and half of the cases in the range of 3. We would therefore be left with x values that are in the range of 3 to 10.

Figure 3 illustrates the second kind of x score distribution that was considered, namely where the x scores are skewed to the right. In this case the "tail" of the distribution is on the right side, and the "hump" is on the left side.

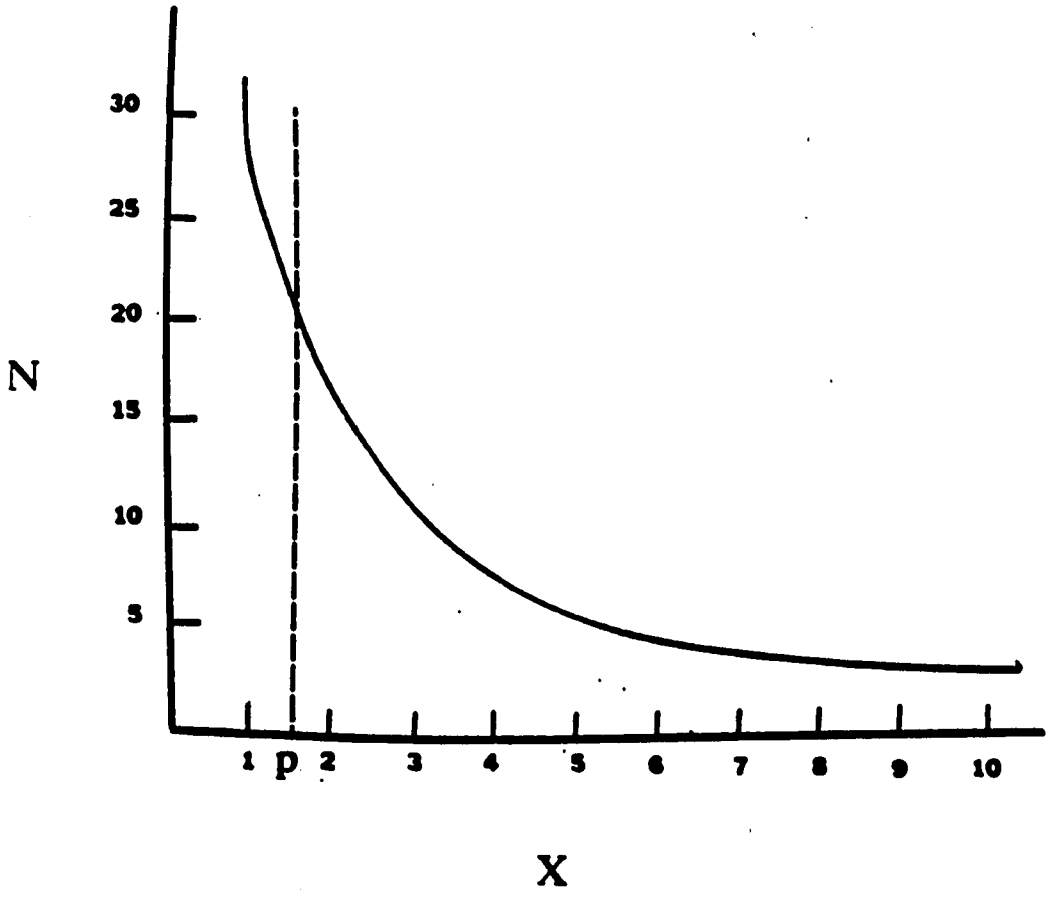


Figure 3. A distribution of x scores that is skewed to the right[a].

[a]p = at a given x value, a proportion, p, is selected.

It should be noted that, in this case, the majority of scores occur at the lower end of the distribution. If one looks at Table 1, it can be seen that 32% of cases occur when x = 1; 20% occur when x = 2; 12% occur when x = 3; 8% occur when x = 4; 8% occur when x = 5; and 4% occur when x = 6, 7, 8, 9 and 10. If we consider the same example described earlier, where N = 100, and the proportion selected is 75% (i.e., all cases to the right of the perpendicular line), we are again deleting the lowest 25 cases. Since there are 32 cases where x = 1, we will still have x values in the range of 1 to 10.

Figure 4 illustrates the third kind of x score distribution that was considered in this study. In this case, the x scores are skewed to the left. It should be noted that the "tail" of the distribution is on the left side, and the "hump" is on the right side.

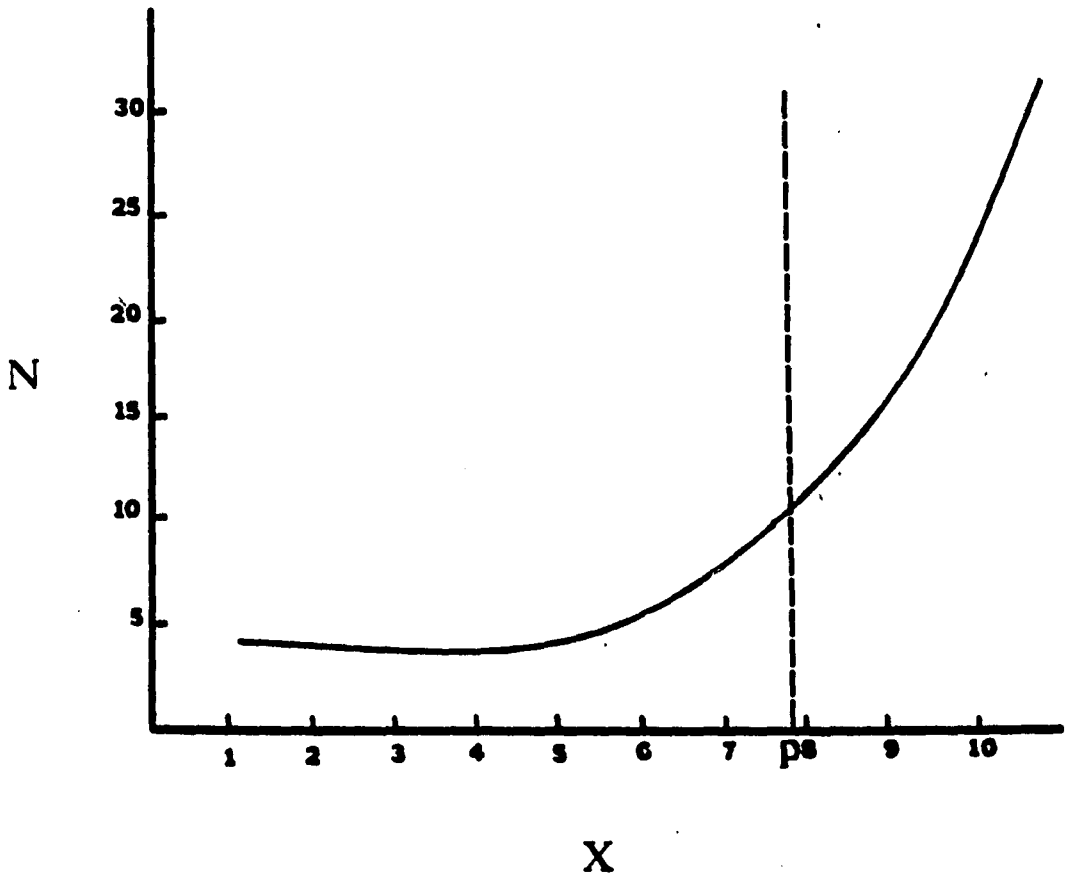


Figure 4. A distribution of x scores that is skewed to the left[a].

[a]p = at a given x value, a proportion, p, is selected.

In this situation, the majority of cases occur at the upper end of the distribution. More specifically, if one looks at Table 1, it is evident that 32% of cases occur when x=10; 20% occur when x = 9; 12% occur when x = 8; 8% occur when x = 7; 8% occur when x = 6; and 4% occur when x = 5, 4, 3, 2 or 1. Let us again consider the example where N = 100, and the proportion selected is 75% (i.e., those cases to the right of the perpendicular line). We again delete the bottom 25 cases. In this case, however, we are left with values for x in the range of 6, 7, 8, 9 and 10 only.

We will now discuss the form of the regression of y on x. The y variable was constructed as a quadratic function of x, and can be described by the following equation:

$$y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + e \qquad (9)$$

Different values for the alpha weights were specified to change the form of the regression of y on x.

Figure 5 provides an illustration of one of the xy relationships. In this particular case, $\alpha_1 = 1$, and $\alpha_2 = 0$. X is in the range of 1 to 10.

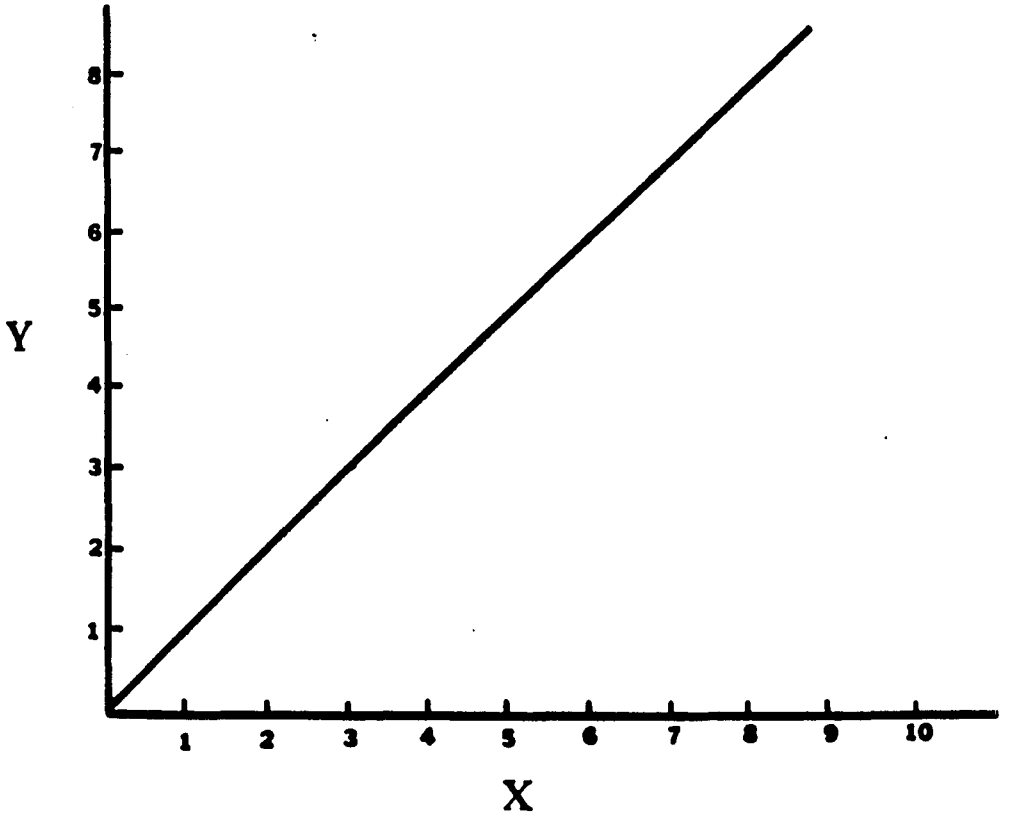


Figure 5. A sample xy relationship that is linear.

It is evident that the relationship between y and x is linear for the particular range of x values utilized here.

Figure 6 provides an illustration of one of the x-y relationships. In this particular case, $\alpha_1 = 1$, and $\alpha_2 = -.05$. X is in the range of 1 to 10.

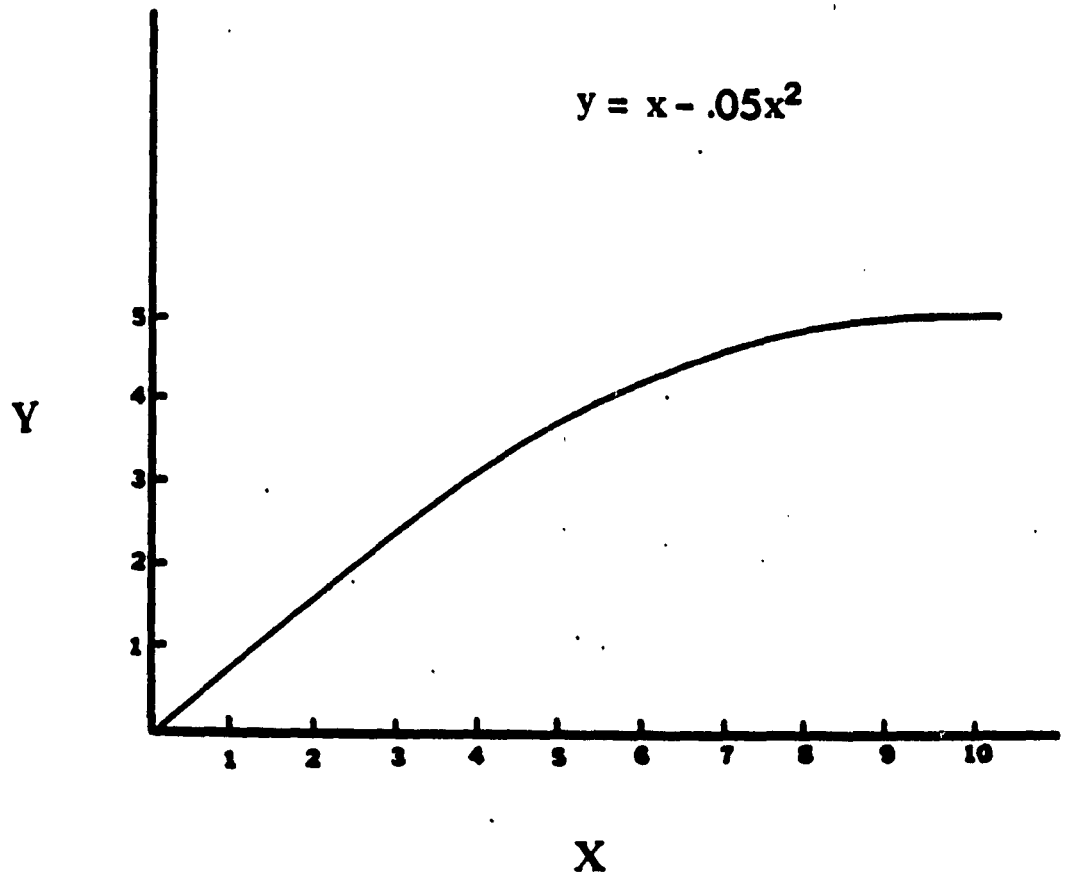


Figure 6. A sample xy relationship that is nonlinear but monotonic.

It is evident that the relationship between y an x is nonlinear but monotonic for the particular range of x values utilized here.

Figure 7 provides an illustration of the x relationship in the case where $\alpha_1 = 5$, and $\alpha_2 = -.5$. x is in the range of 1 to 10.

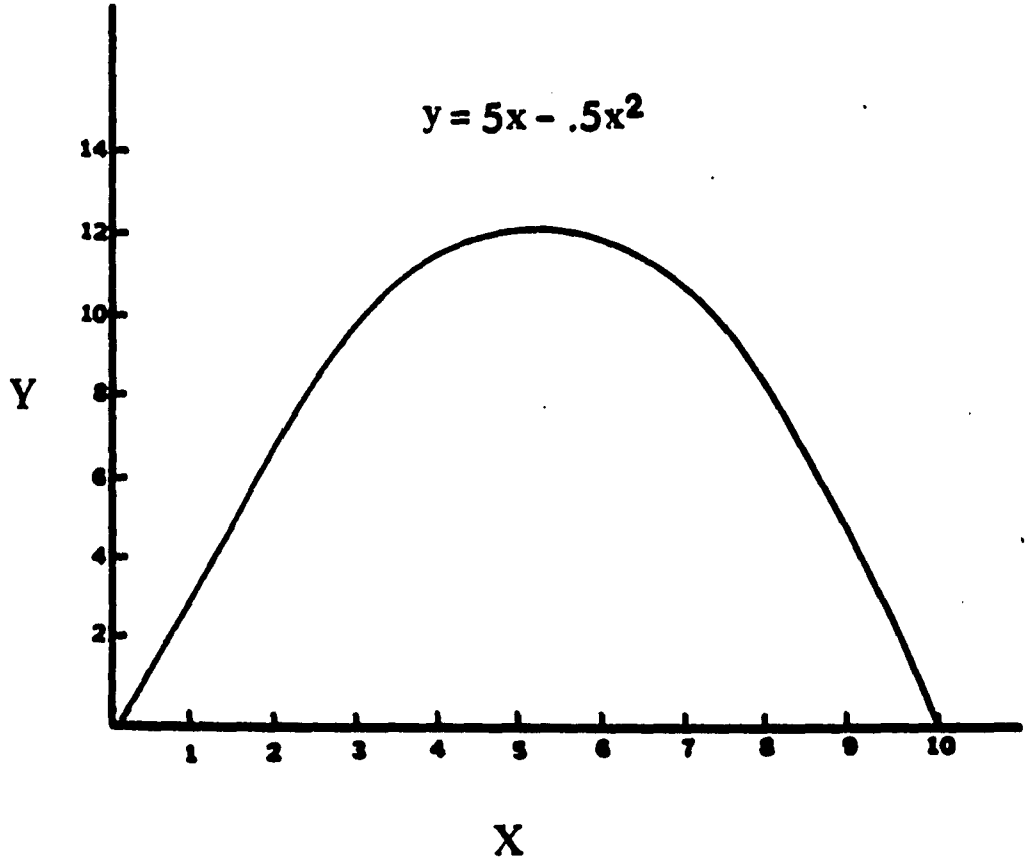


Figure 7. A sample xy relationship that is nonlinear and non-monotonic.

In this case, the x relationship is nonlinear and non-monotonic for x scores in the range of 1 to 10.

The form of the regression of z on x will now be elaborated upon. A FORTRAN computer program was utilized to generate a z variable from a Guassian distribution, where a random number was utilized to generate sample z values. Z scores were generated that were related to x, but not as highly related to x as is x^2. More specifically, z was assumed to be normally distributed with mean, $w_0 + w_1x + w_2x^2$, and the variance was set equal to 1. Values for z were then generated from this distribution. Figure 8 illustrates the two different xz relationships that were evaluated in the present study.

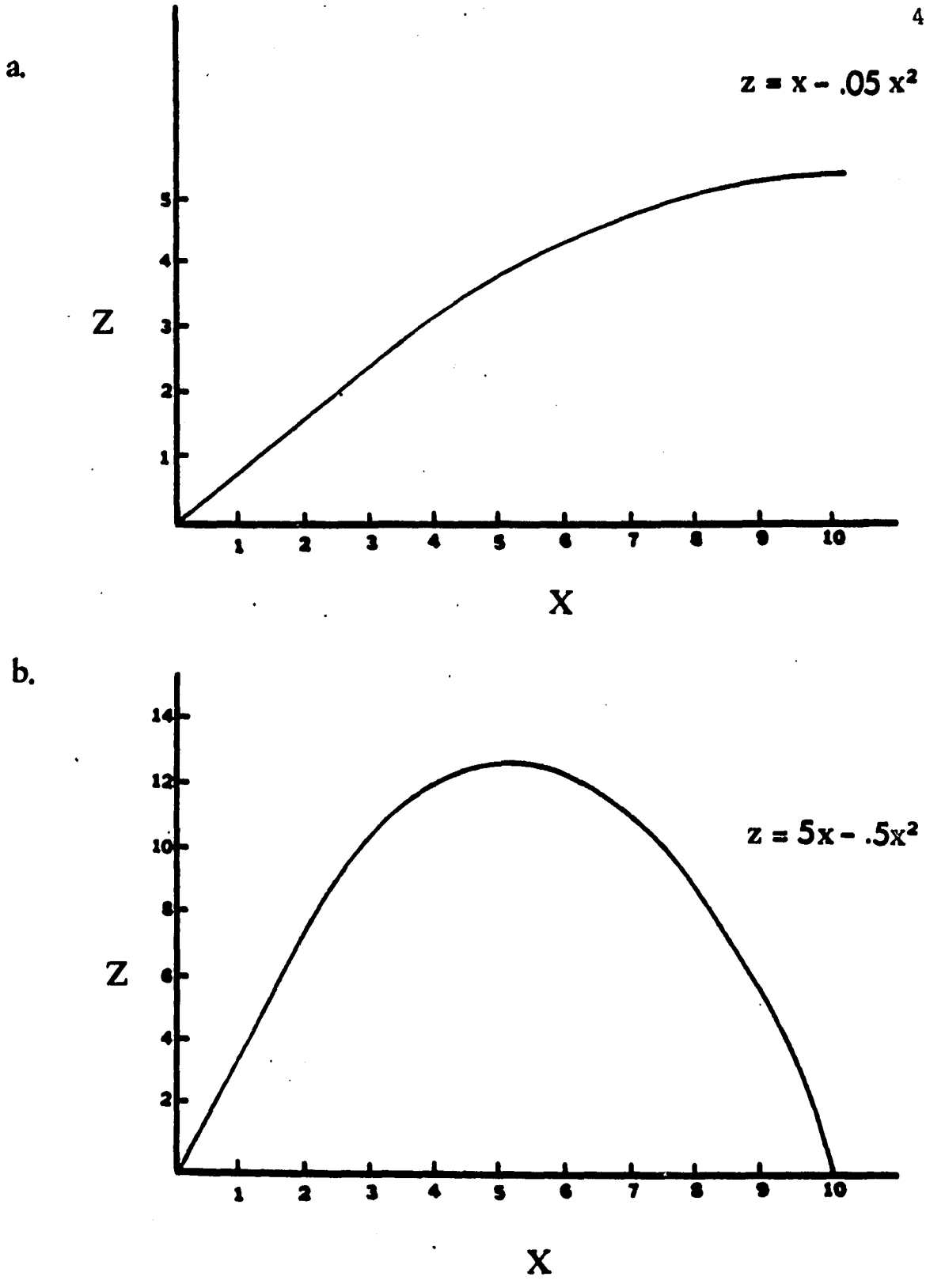


Figure 8. Two sample xy relationships: a. nonlinear but monotonic; b. nonlinear and non-monotonic.

In Figure 8 (a) and (b), the relationship of z to x can be described by the following expression:

$$E(z/x) = w_0 + w_1(x) + w_2(x)^2 \qquad (10)$$

where w_0, w_1, and w_2 refer to the regression coefficients of z_0, z_1, and z^2 respectively, in the regression equation predicting x from z. In Figure 8(a), the relationship of z to x is nonlinear but monotonic for x value that range between 1 and 10, and can be described by the following equation: $x = z - .05(z)^2$. (The variance of z given x was specified to be 1). In figure 8(b), the relationship of z to x is nonlinear and non-monotonic for the range of x values between 1 and 10, and is described by the following equation: $x = 5(z) - .5(z)^2$. (The variance of z given x was specified to be 1).

One can conceive of procedure 3 (where $z = x^2$) as a special case of procedure 2. The relationship of z to x is illustrated by Figure 9.

In this case, the relationship of z to x is nonlinear but monotonic, and can be expressed by the prototype illustrated by equation 8 above. In this particular case the values for w_0, w_1, and w_2 are 0, 0, and 1 respectively. (The variance of z given x is 0).

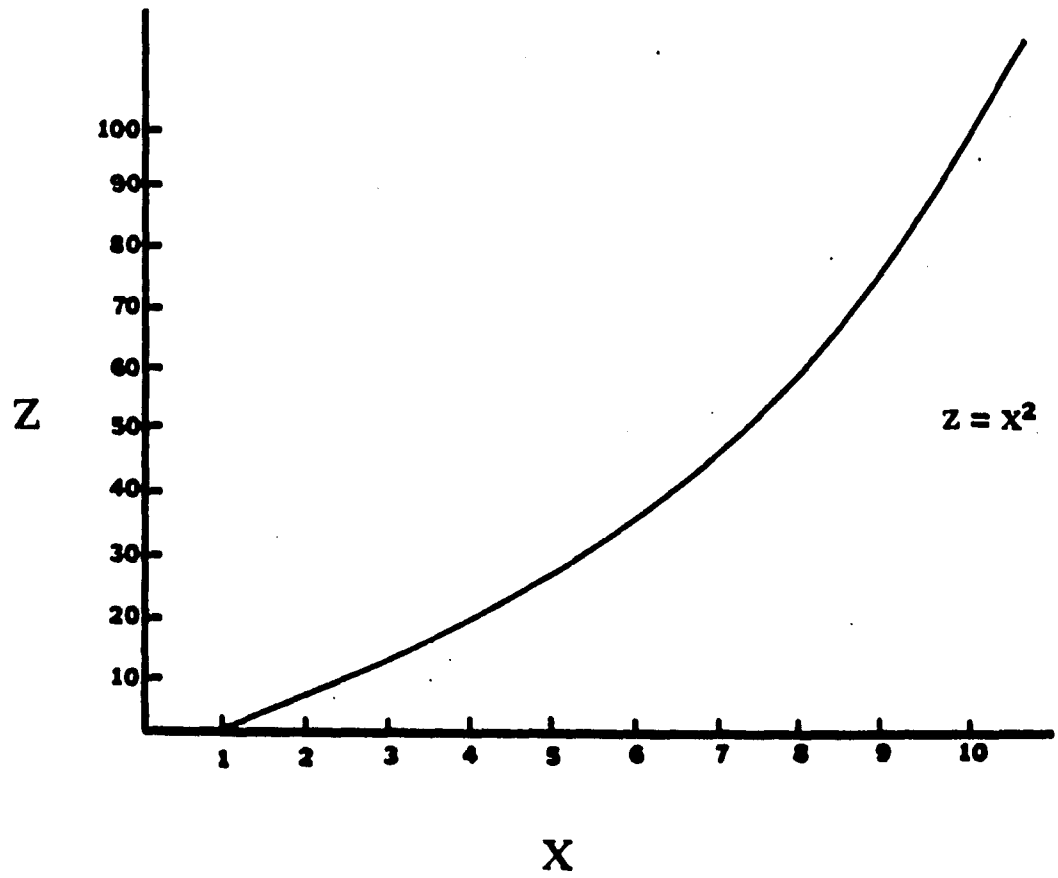


Figure 9. A sample quadratic relationship between x and z.

Chapter IV

RESULTS

Tables 2 through 10 represent the effects of the five independent variables on the performance of indices 1, 2, and 3. Index 1 measures the accuracy of predicting missing Y scores, utilizing a linear regression equation of the selected cases only (i.e., evaluating procedure 1). Index 2 measures the accuracy of predicting missing y scores, where a variable, z, is used in conjunction with x in a linear regression analysis (i.e., evaluating procedure 2). Index 3 measures the accuracy of predicting missing y scores, where a variable, x^2, is used in conjunction with x in a linear regression analysis (i.e., evaluating Procedure 3).

In each of tables 2 through 10, for a given value of N (number of cases) and PSEL (proportion selected), the performance of the three indices is presented as a function of XDIST (x distribution is: 1. uniform; 2. skewed to the right; 3. skewed to the left); ALPHA (the relationship of y to x is: 1. nonlinear but monotonic; 2. nonlinear and non-monotonic; 3. linear); W (the relationship of z to x is: 1. nonlinear but monotonic, 2. nonlinear and non-monotonic). For example, Table 2

demonstrates the performance of the three indices for each category of alpha, (α), W, and Xdist, where N and Psel are fixed at 25 and 60% respectively. Thus, tables 2 through 10 represent all the basic results.

The following are the major findings observed in tables 2 through 10, which describe the main results of the study:

1. When the form of the regression of y on x was linear, index 1 was, as expected, always smaller than indices 2 and 3. Therefore, in this case, the best procedure was to predict missing y scores using a simple linear regression equation predicting y from the selected x cases.

2. When the form of the regression of y on x was monotonic, the values for indices 1 and 2 were similar. The values for index 3 fluctuated more widely than those for indices 1 and 2. Values for index 3 became extremely large in the case where the distribution of x scores was skewed to the left.

3. In the case where the form of the regression of y on x was non-monotonic, the performance of procedure 3 was superior to that of methods 1 and 2, as evidenced by the much smaller values for index 3.

4. The performance of index 3 remained constant across all levels of alpha.

Table 2

The Values for Indices 1, 2, and 3 as Functions of Alpha (α), W, and Type of x Distribution (n = 25; Psel = .60)

Form of the regression of y on x											
Linear				Monotonic				Non-Monotonic			
		Form of the regression of z on x				Form of the regression of z on x				Form of the regression of z on x	
Distribution of x scores	Index	Monotonic	Non-Monotonic	Distribution of x scores	Index	Monotonic	Non-Monotonic	Distribution of x scores	Index	Monotonic	Non-Monotonic
Uniform	1	1.63	1.63	Uniform	1	3.32	3.32	Uniform	1	170.85	170.85
Uniform	2	1.74	4.51	Uniform	2	3.33	4.85	Uniform	2	160.99	38.92
Uniform	3	8.49	8.49	Uniform	3	8.49	8.49	Uniform	3	8.48	8.48
Right-skewed	1	1.24	1.24	Right-skewed	1	1.76	1.76	Right-skewed	1	53.5	53.5
Right-skewed	2	1.37	1.65	Right-skewed	2	1.87	1.67	Right-skewed	2	51.54	3.85
Right-skewed	3	1.73	1.73	Right-skewed	3	1.73	1.73	Right-skewed	3	1.73	1.73
Left-skewed	1	5.41	5.41	Left-skewed	1	7.67	7.67	Left-skewed	1	231.07	231.07
Left-skewed	2	5.85	30.75	Left-skewed	2	8.15	33.2	Left-skewed	2	235.55	271.67
Left-skewed	3	424.44	424.44	Left-skewed	3	424.42	424.42	Left-skewed	3	424.49	424.49

Note. Numerical values represent the expected mean squared error of prediction for the three procedures.

Table 3

The Values for Indices 1, 2, and 3 as Functions of Alpha, N, and Type of x Distribution (n = 25; rho = .75)

Form of the regression of y on x											
Linear				Monotonic				Non-Monotonic			
		Form of the regression of z on x				Form of the regression of z on x				Form of the regression of z on x	
Distribution of x scores	Index	Monotonic	Non-Monotonic	Distribution of x scores	Index	Monotonic	Non-Monotonic	Distribution of x scores	Index	Monotonic	Non-Monotonic
Uniform	1	1.3	1.3	Uniform	1	2.44	2.44	Uniform	1	115.92	115.92
Uniform	2	1.48	2.51	Uniform	2	2.36	2.53	Uniform	2	109.08	5.3
Uniform	3	2.54	2.54	Uniform	3	2.54	2.54	Uniform	3	2.54	2.54
Right-skewed	1	1.15	1.15	Right-skewed	1	1.46	1.46	Right-skewed	1	32.2	32.2
Right-skewed	2	1.29	1.32	Right-skewed	2	1.59	1.34	Right-skewed	2	30.83	3.25
Right-skewed	3	1.33	1.33	Right-skewed	3	1.33	1.33	Right-skewed	3	1.33	1.33
Left-skewed	1	2.65	2.65	Left-skewed	1	5.1	5.1	Left-skewed	1	247.35	247.35
Left-skewed	2	3.29	23.91	Left-skewed	2	5.8	25.53	Left-skewed	2	255.63	186.6
Left-skewed	3	65.77	65.77	Left-skewed	3	65.79	65.79	Left-skewed	3	65.79	65.79

Note. Numerical values represent the expected mean squared error of prediction for the three procedures.

Table 4

The Values for Indices 1, 2, and 3 as Functions of Alpha (α), N, and Type of x Distribution (n = 25; Psel = .10)

Form of the regression of y on x

Linear				Monotonic				Non-Monotonic			
		Form of the regression of z on x				Form of the regression of z on x				Form of the regression of z on x	
Distribution of x scores	Index	Monotonic	Non-Monotonic	Distribution of x scores	Index	Monotonic	Non-Monotonic	Distribution of x scores	Index	Monotonic	Non-Monotonic
Uniform	1	1.19	1.19	Uniform	1	2.0	2.0	Uniform	1	82.67	82.67
Uniform	2	1.29	1.55	Uniform	2	1.96	1.57	Uniform	2	68.16	3.79
Uniform	3	1.58	1.58	Uniform	3	1.58	1.58	Uniform	3	1.58	1.58
Right-skewed	1	1.09	1.09	Right-skewed	1	1.22	1.22	Right-skewed	1	13.43	13.43
Right-skewed	2	1.15	1.16	Right-skewed	2	1.25	1.17	Right-skewed	2	10.81	1.77
Right-skewed	3	1.14	1.14	Right-skewed	3	1.14	1.14	Right-skewed	3	1.14	1.14
Left-skewed	1	1.64	1.64	Left-skewed	1	3.71	3.71	Left-skewed	1	208.3	208.3
Left-skewed	2	2.23	6.61	Left-skewed	2	4.51	6.64	Left-skewed	2	230.96	9.75
Left-skewed	3	5.67	5.67	Left-skewed	3	5.67	5.67	Left-skewed	3	5.67	5.67

Note. Numerical values represent the expected mean squared error of prediction for each of the three procedures.

Table 5

The Values for Indices 1, 2, and 3 as Functions of Alpha (α), N, and Type of x Distribution (n = 50; Psel = .60)

Form of the regression of y on x											
Linear				Monotonic				Non-Monotonic			
		Form of the regression of z on x				Form of the regression of z on x				Form of the regression of z on x	
Distribution of x scores	Index	Monotonic	Non-Monotonic	Distribution of x scores	Index	Monotonic	Non-Monotonic	Distribution of x scores	Index	Monotonic	Non-Monotonic
Uniform	1	1.35	1.35	Uniform	1	3.03	3.03	Uniform	1	169.13	169.13
Uniform	2	1.4	3.84	Uniform	2	3.08	4.14	Uniform	2	169.09	33.85
Uniform	3	5.21	5.21	Uniform	3	5.21	5.21	Uniform	3	5.18	5.18
Right-skewed	1	1.11	1.11	Right-skewed	1	1.6	1.6	Right-skewed	1	49.99	49.99
Right skewed	2	1.15	1.26	Right-skewed	2	1.64	1.3	Right-skewed	2	50.52	4.36
Right skewed	3	1.33	1.33	Right-skewed	3	1.33	1.33	Right-skewed	3	1.33	1.33
Left-skewed	1	2.89	2.89	Left-skewed	1	5.17	5.17	Left-skewed	1	230.71	230.71
Left-skewed	2	2.94	9.17	Left-skewed	2	5.22	11.51	Left-skewed	2	230.33	242.38
Left-skewed	3	214.25	214.25	Left-skewed	3	213.83	213.83	Left-skewed	3	213.82	213.82

Note. Numerical values represent the expected mean squared error of prediction for each of the three procedures.

Table 6

The Values for Indices 1, 2, and 3 as Functions of Alpha (α), W, and Type of x Distribution (n = 50; Psel = .75)

Form of the regression of y on x

Linear				Monotonic				Non-Monotonic			
		Form of the regression of z on x				Form of the regression of z on x				Form of the regression of z on x	
Distribution of x scores	Index	Monotonic	Non-Monotonic	Distribution of x scores	Index	Monotonic	Non-Monotonic	Distribution of x scores	Index	Monotonic	Non-Monotonic
Uniform	1	1.17	1.17	Uniform	1	2.3	2.3	Uniform	1	113.48	113.48
Uniform	2	1.21	1.69	Uniform	2	2.33	1.76	Uniform	2	113.86	8.95
Uniform	3	1.87	1.87	Uniform	3	1.87	1.87	Uniform	3	1.87	1.87
Right-skewed	1	1.07	1.07	Right-skewed	1	1.34	1.34	Right-skewed	1	28.02	28.02
Right-skewed	2	1.1	1.12	Right-skewed	2	1.41	1.14	Right-skewed	2	31.82	3.48
Right-skewed	3	1.14	1.14	Right-skewed	3	1.14	1.14	Right-skewed	3	1.13	1.13
Left-skewed	1	1.68	1.68	Left-skewed	1	3.97	3.97	Left-skewed	1	230.60	230.60
Left-skewed	2	1.72	5.58	Left-skewed	2	4.01	6.91	Left-skewed	2	230.74	138.93
Left-skewed	3	17.35	17.35	Left-skewed	3	17.35	17.35	Left-skewed	3	17.32	17.32

Note. Numerical values represent the expected mean squared error of prediction for each of the three procedures.

Table 7

The Values for Indices 1, 2, and 3 as Functions of Alpha (α), N, and Type of x Distribution (n = 50; Psel = .10)

Form of the regression of y on x											
Linear				Monotonic				Non-Monotonic			
		Form of the regression of z on x				Form of the regression of z on x				Form of the regression of z on x	
Distribution of x scores	Index	Monotonic	Non-Monotonic	Distribution of x scores	Index	Monotonic	Non-Monotonic	Distribution of x scores	Index	Monotonic	Non-Monotonic
Uniform	1	1.11	1.11	Uniform	1	1.92	1.92	Uniform	1	82.38	82.38
Uniform	2	1.16	1.35	Uniform	2	1.99	1.37	Uniform	2	84.39	3.41
Uniform	3	1.33	1.33	Uniform	3	1.33	1.33	Uniform	3	1.31	1.31
Right-skewed	1	1.05	1.05	Right-skewed	1	1.17	1.17	Right-skewed	1	13.37	13.37
Right-skewed	2	1.09	1.07	Right-skewed	2	1.22	1.09	Right-skewed	2	13.9	3.01
Right-skewed	3	1.07	1.07	Right-skewed	3	1.07	1.07	Right-skewed	3	1.07	1.07
Left-skewed	1	1.32	1.32	Left-skewed	1	3.39	3.39	Left-skewed	1	207.97	207.97
Left-skewed	2	1.42	3.22	Left-skewed	2	3.57	3.34	Left-skewed	2	216.46	16.14
Left-skewed	3	3.34	3.34	Left-skewed	3	3.33	3.33	Left-skewed	3	3.32	3.32

Note. Numerical values represent the expected mean squared error of prediction for each of the three procedures.

Table 8

The Values for Indices 1, 2, and 3 as Functions of Alpha (α), W, and Type of x Distribution (n = 100; rse = .60)

Form of the regression of y on x											
Linear				Monotonic				Non-Monotonic			
		Form of the regression of z on x				Form of the regression of z on x				Form of the regression of z on x	
Distribution of x scores	Index	Monotonic	Non-Monotonic	Distribution of x scores	Index	Monotonic	Non-Monotonic	Distribution of x scores	Index	Monotonic	Non-Monotonic
Uniform	1	1.16	1.16	Uniform	1	2.84	2.84	Uniform	1	168.56	168.56
Uniform	2	1.24	2.58	Uniform	2	2.94	2.81	Uniform	2	170.92	25.68
Uniform	3	3.03	3.03	Uniform	3	3.02	3.02	Uniform	3	2.98	2.98
Right-skewed	1	1.05	1.05	Right-skewed	1	1.53	1.53	Right-skewed	1	48.53	48.53
Right skewed	2	1.09	1.16	Right-skewed	2	1.48	1.17	Right-skewed	2	39.78	2.19
Right skewed	3	1.16	1.16	Right skewed	3	1.16	1.16	Right-skewed	3	1.15	1.15
Left-skewed	1	1.86	1.86	Left-skewed	1	4.19	4.19	Left-skewed	1	232.17	232.17
Left skewed	2	2.03	6.56	Left-skewed	2	4.4	8.88	Left-skewed	2	235.61	235.34
Left-skewed	3	90.2	90.2	Left-skewed	3	89.57	89.57	Left-skewed	3	89.22	89.22

Note. Numerical values represent the expected mean squared error of prediction for each of the three procedures.

Table 9

The Values for Indices 1, 2, and 3 as Functions of Alpha (α), W, and Type of x Distribution (n = 100; Psel = .15)

Form of the regression of y on x											
Linear				Monotonic				Non-Monotonic			
		Form of the regression of z on x				Form of the regression of z on x				Form of the regression of z on x	
Distribution of x scores	Index	Monotonic	Non-Monotonic	Distribution of x scores	Index	Monotonic	Non-Monotonic	Distribution of x scores	Index	Monotonic	Non-Monotonic
Uniform	1	1.08	1.08	Uniform	1	2.17	2.17	Uniform	1	110.39	110.39
Uniform	2	1.09	1.3	Uniform	2	2.09	1.38	Uniform	2	100.94	8.86
Uniform	3	1.4	1.4	Uniform	3	1.4	1.40	Uniform	3	1.4	1.4
Right skewed	1	1.03	1.03	Right-skewed	1	1.28	1.28	Right-skewed	1	26.23	26.23
Right skewed	2	1.04	1.06	Right-skewed	2	1.26	1.07	Right-skewed	2	22.97	2.26
Right skewed	3	1.07	1.07	Right-skewed	3	1.07	1.07	Right-skewed	3	1.06	1.06
Left skewed	1	1.29	1.29	Left-skewed	1	3.57	3.57	Left-skewed	1	227.38	227.38
Left skewed	2	1.32	2.84	Left skewed	2	3.54	3.93	Left-skewed	2	223.58	111.82
Left skewed	3	7.85	7.85	Left-skewed	3	7.86	7.86	Left-skewed	3	7.78	7.78

Note. Numerical values represent the expected mean squared error of prediction for each of the three procedures.

Table 10

The Values for Indices 1, 2, and 3 as Functions of Alpha (α), W, and Type of x Distribution (n = 100; Rsel = .70)

Form of the regression of y on x											
Linear				Monotonic				Non-Monotonic			
		Form of the regression of z on x				Form of the regression of z on x				Form of the regression of z on x	
Distribution of x scores	Index	Monotonic	Non-Monotonic	Distribution of x scores	Index	Monotonic	Non-Monotonic	Distribution of x scores	Index	Monotonic	Non-Monotonic
Uniform	1	1.05	1.05	Uniform	1	1.87	1.87	Uniform	1	83.16	83.16
Uniform	2	1.06	1.14	Uniform	2	1.83	1.17	Uniform	2	77.17	4.75
Uniform	3	1.16	1.16	Uniform	3	1.16	1.16	Uniform	3	1.11	1.11
Right-skewed	1	1.02	1.02	Right-skewed	1	1.14	1.14	Right-skewed	1	12.77	12.77
Right-skewed	2	1.03	1.03	Right-skewed	2	1.17	1.05	Right-skewed	2	14.42	2.47
Right-skewed	3	1.03	1.03	Right-skewed	3	1.03	1.03	Right-skewed	3	1.0	1.0
Left-skewed	1	1.14	1.14	Left-skewed	1	3.05	3.05	Left-skewed	1	191.16	191.16
Left-skewed	2	1.16	1.63	Left-skewed	2	3.0	1.8	Left-skewed	2	184.76	18.79
Left-skewed	3	1.86	1.86	Left-skewed	3	1.86	1.86	Left-skewed	3	1.84	1.84

Note. Numerical values represent the expected mean squared error of prediction for each of the three procedures.

It should be noted that when the distribution of x scores was uniform, and the number of cases was 30, the values of the three indices should actually be slightly larger in order to be comparable with the right-skewed and left-skewed distributions. This is due to the fact that these latter distributions were based upon an n of 25.

Tables 11 through 15 summarize tables 2 through 10 in terms of the main effects of each of the five independent variables on the values of the three indices. These tables (and the subsequent tables indicating interaction effects) were obtained by performing an analysis of variance on the data, where the dependent variable was the EMSE values, and each of the five independent variables was considered as a separate factor. Each main effect was derived by collapsing over values of the other variables. The following major findings were observed:

Table 11: The main effect of N.

1. As the number of cases (n) increased, the value of each of the three indices decreased, indicating a better prediction system for the missing Y scores.
2. The values for index 2 were lower than those for index 1 at each level of n.
3. The value of index 3 was larger than indices 1 and 2 when n was 30, but was smaller than indices 1 and 2 when

Table 11

Average Values for Indices 1, 2, and 3
as Functions of the Number of Cases (N)

	N		
Index	25[a]	50	100
1	44.49	43.05	41.95
2	34.84	31.37	28.85
3	56.97	27.4	12.02

25[a]. In the case of a uniform distribution, n = 30.

Table 12

Values for Indices 1, 2, and 3 as Functions of the Proportion of Cases Selected (Psel)

	Psel		
Index	.60	.75	.90
1	51.91	43.24	34.27
2	44.16	31.74	19.17
3	83.22	11.15	2.01

Table 13

Values for Indices 1, 2, and 3 as Functions of the xy Relationship (α)

	xy Relationship (α)		
Index	Linear	Monotonic	Non-Monotonic
1	1.51	2.75	125.23
2	3.07	3.91	88.08
3	32.2	32.2	32.2

Table 14

Values for Indices 1, 2, and 3 as Functions of the xz Relationship (W)

	xz Relationship (W)	
Index	Monotonic	Non-Monotonic
1	43.16	43.16
2	43.03	20.34
3	32.13	32.13

Table 15

Values for Indices 1, 2, and 3 as Functions of the Distribution of x Scores

	x Distribution		
Index	Uniform	Right-skewed	Left-skewed
1	41.83	11.12	76.53
2	23.4	6.27	65.39
3	2.95	1.22	92.21

n was 50 and 100. Further, the values for index 3 fluctuated widely over n in comparison to indices 1 and 2.

Table 12: The main effect of PSEL.

1. As the proportion selected (psel) increased, the value of each of the three indices decreased.
2. The values for index 2 were lower than those of index 1 at each level of psel.
3. The value of index 3 was larger than indices 1 and 2 when psel was .60, but was much smaller than those of indices 1 and 2 when psel was .75 and .90. Further, the values for index 3 fluctuated widely over psel in comparison to indices 1 and 2.

Table 13: The main effect of the xy relationship (α). 1. As alpha (which determines the relationship between x and y) changed from nonlinear but monotonic to nonlinear and non-monotonic, the values for indices 1 and 2 increased.
2. The values for indices 1 and 2 were much higher when the xy relationship was non-monotonic.
3. The values for index 3 remained constant at each level of alpha.

Table 14: The main effect of the xz relationship ($\underline{W}$).

1. As the relationship between z and x changed from nonlinear but monotonic to nonlinear and non-monotonic,

the values for index 2 decreased. In the case where the xy relationship was non-monotonic, the values for index 2 were substantially smaller than those of indices 1, and 3, indicating a better prediction system with the utilization of index 2.

2. There was no difference in the values of indices 1 and 3 when the x-z relationship was varied. This was expected, as the calculation of indices 1 and 3 do not depend upon W.

Table 15: the main effect of x distribution.

1. As the distribution of x scores changed from skewed to the right to uniform, to skewed to the left, the values of all indices increased correspondingly.

2. In the case where the distribution of x scores was skewed to the left, there was the greatest amount of increase in the values for all indices. This effect was especially apparent for index 3.

3. At each type of X score distribution, the values for index 2 were smaller than those for index 1.

4. The values for index 3 were smaller than those for indices 1 and 2, in the cases where the distribution of x scores were uniform and skewed to the right. However, in the case where the distribution of x scores was skewed to the left, the value for index 3 was larger than those of indices 1 and 2.

Tables 16 through 20 represent the 2-way interaction effects among several of the variables that were derived from the analysis of variance design. It should be noted that most of the significant interactions included the distribution of X scores as one of the variables.

Table 16: The interaction of XDIST and N.

1. All three indices were most sensitive to the form of the X distribution when sample sizes were small. The procedure which used x and x^2 to predict missing Y scores (represented by index 3) was especially poor when the distribution of x scores was skewed to the left, and the sample size was small.
2. When the number of cases increased from 25 to 50 and 100, there was a dramatic improvement in procedure 3 (as evidenced by a decrease in the values for index 3). When the number of cases was 100, the values for index 3 were smaller than those of indices 1 and 2.
3. The values for index 2 were always smaller than those of index 1, at each condition of n and distribution of x scores.

Table 17: The interaction of XDIST and PSEL.

1. All three indices were most sensitive to the type of x distribution when the proportion of selection was .60. The performance of index 3 was especially poor in the

Table 16

Average Values for Indices 1, 2, and 3 Broken Down by the Distribution of x Scores and the Number of Cases

X Distribution	Index 1	Index 2	Index 3
n = 25[a]			
Uniform	42.37	23.1	4.2
Right-skewed	11.89	6.6	1.4
Left-skewed	79.21	74.81	165.3
n = 50			
Uniform	41.76	24.38	2.8
Right-skewed	10.97	6.7	1.18
Left-skewed	76.41	62.98	78.22
n = 100			
Uniform	41.36	22.72	1.8
Right-skewed	10.51	5.4	1.08
Left-skewed	73.97	58.38	33.12

25[a]. In the case of a uniform distribution, n = 30.

Table 17

Values for Indices 1, 2, and 3 Broken Down by the Distribution of x Scores and by the Proportion of Cases Selected

x Distribution	Index 1	Index 2	Index 3
Proportion Selected = .60			
Uniform	57.99	35.33	5.57
Right-skewed	17.81	9.39	1.41
Left-skewed	80.13	87.75	242.69
Proportion Selected = .75			
Uniform	38.92	20.48	1.94
Right-skewed	10.42	6.08	1.18
Left-skewed	80.39	68.65	30.32
Proportion Selected = .90			
Uniform	28.59	14.39	1.35
Right-skewed	5.14	3.33	1.08
Left-skewed	69.08	39.78	3.62

Table 18

Values for Indices 1, 2, and 3 Broken Down by the Distribution of x Scores and the xy Relationship (α)

x Distribution	Index 1	Index 2	Index 3
xy Relationship: Linear			
Uniform	1.23	1.79	2.96
Right-skewed	1.09	1.17	1.22
Left-skewed	2.21	6.23	92.31
xy Relationship: Monotonic			
Uniform	2.43	2.42	2.96
Right-skewed	1.39	1.33	1.22
Left-skewed	4.42	8.0	92.19
xy Relationship: Non-Monotonic			
Uniform	121.89	66.01	2.98
Right-skewed	30.89	16.3	1.22
Left-skewed	222.96	181.95	92.14

Table 19

Values for Indices 1, 2, and 3 Broken Down by the Proportion Selected (Psel) and the Number of Cases (N)

Psel	Index 1	Index 2	Index 3
	n = 25		
.60	52.94	47.86	144.89
.75	45.51	36.87	23.22
.90	35.03	19.8	2.8
	n = 50		
.60	51.66	43.18	73.5
.75	42.63	30.9	6.79
.90	34.85	19.96	1.91
	n = 100		
.60	51.32	41.44	31.28
.75	41.59	27.36	3.43
.90	32.93	17.75	1.34

Table 20

Values for Indices 1, 2, and 3 Broken Down by the Proportion Selected (Psel) and the xy Relationship (α)

Psel	Index 1	Index 2	Index 3
	xy Relationship: Linear		
.60	1.97	4.46	83.32
.75	1.38	3.05	11.15
.90	1.18	1.69	2.02
	xy Relationship: Monotonic		
.60	3.46	5.65	83.32
.75	2.63	3.89	11.15
.90	2.16	2.21	2.02
	xy Relationship: Non-Monotonic		
.60	150.5	122.36	83.32
.75	125.72	88.28	11.15
.90	99.47	53.61	2.02

case where the distribution of x scores was skewed to the left (and psel was .60).

2. When the proportion selected increased from 60% to 75% and 90%, there was a dramatic improvement in the performance of procedure 3, i.e., the values for index 3 were smaller than those of indices 1 and 2 in the case where the distribution of x scores was skewed to the left.

3. At each condition of psel, and type of x distribution, the values for index 2 were always smaller than those of index 1.

Table 18: The interaction of XDIST and the xy relationship.

1. The effect of distribution of X scores was greatest when the relationship between X and Y was non-monotonic. The interaction was observed in the situation where the distribution of X scores was skewed to the left.

2. The values for index 3 remained constant across all levels of alpha.

Table 19: The interaction of the proportion selected (PSEL) and the number of cases (N)_.

1. This interaction effect is significant for index 3 alone, where the greatest change in index values occurred in the cases where the proportion selected went from 60% to 75%.

Table 20: The interaction of proportion selected (PSEL) and the xy relationship (alpha).

1. The interaction effect is significant for indices 1 and 2, in the case where the relationship between x and y is non-monotonic.

Table 21 represents the mean squared error values for indices 1,2 and 3, and was derived from the analysis of variance design by collapsing over all the independent variables.

Table 21

Values for Indices 1, 2, and 3, Collapsing Over all Independent Variables

Index		
1	2	3
43.16	31.69	32.13

It is evident that indices 2 and 3 are best, and index 1 is worst (when collapsed over all independent variables).

It should be noted that the ordering of the three indices in this case, is largely due to the fact that

the values for index 3 were much smaller than those for indices 1 and 2 when the regression of y on x was non-monotonic. For example, Table 2 demonstrates a case where the xy relationship is non-monotonic, the distribution of x scores is uniform, and the xz relationship is monotonic. In this case the values for indices 1, 2, and 3 were 170.85, 160.99, and 8.48 respectively.

In summary, procedure 1, where missing y scores were estimated using a linear regression equation predicting y from the selected x cases, produced, as expected, a better estimate than the other two procedures when the underlying xy relationship was linear. Procedure 2, which used an additional variable, z, in conjunction with x to predict missing y scores, produced a better estimate than procedure 1 when the xy relationship was non-monotonic. In this case, the values for index 2 were much lower than those of index 1 when the xz relationship was non-monotonic. Procedure 3, which used x^2 in conjunction with x to predict missing y scores, produced the best estimate in the case where the relationship among x and y was non-monotonic. The values for index 3 were much lower, in this case, than those for indices 1 and 2. However, procedure 3 produced the poorest estimates (and the values for index 3 were quite large in comparison with indices 1 and 2)

when the distribution of x scores was skewed to the left.

Chapter V

SUMMARY AND DISCUSSION

This study investigated the problem of analyzing results with missing data. More specifically, we focused on a situation involving two variables, x and y, where the relationship between x and y in the population was nonlinear in form. Data on y were missing as a function of x. We were interested in estimating the missing y scores.

Three procedures for estimating the missing scores were considered. In the first procedure, a linear least squares analysis, where y was regressed on x, was applied to the subset of selected cases. The resulting equation was then utilized to estimate the missing y scores. This procedure was called the "selected cases" method.

The second procedure introduced an additional variable, z, where complete data on z were available. The z variable served as a replacement for x^2. The y variable was regressed on x and z in a linear least squares analysis, using x, z, and y data for the subset of selected cases. The resulting equation was applied to x and z in the unselected group to predict missing y

scores. This procedure was called the "reproduced cases" method.

The third procedure was a special case of the second procedure, where the variable z was set to be x^2. Thus, y was regressed on X and X^2 using x, x^2, and y data for the subset of selected cases. The resulting equation was applied to x and x^2 in the unselected group to predict missing y scores.

The accuracy of the three procedures was considered in terms of how well each procedure could reproduce the missing scores. Expressions for the expected mean squared error (EMSE) of each procedure were analytically derived (see equation 7).

One major finding was that the performance of procedure 1 was often superior to that of the other two procedures, i.e., having the smallest EMSE overall, and always superior in the situation where the relationship between x and y was linear.

Of the three procedures considered, one might have expected that the performance of procedure 3 would be reasonably good, as it utilizes the correct model. That is, since the correct model specifies y as a function of x and x^2, the utilization of x^2 together with x (in procedure 3) should make for a good prediction system. However, results demonstrated that the performance of procedure 3 was often poorer than that of procedures 1

and 2, and especially poor when the distribution of x scores was skewed to the left. This was demonstrated by the larger values for EMSE, (with the exception of the case where the form of the regression of y on x was non-monotonic). A possible explanation for this result can be found if we examine this method further. In the first step of this procedure we are using the values of two highly correlated variables, x and x^2, in the selected group, to predict the values of y in the selected group (utilizing a linear regression analysis). The high correlation between x and x^2 can lead to large standard errors of the obtained regression weights for the selected sample. This high intercorrelation among x and x^2 is especially evident in the case where the distribution of x scores is skewed to the left, and selection based upon x severely restricts the range of x. However, this should not necessarily lead to large EMSE values (i.e., poor estimation for the predicted Y scores), since different values for the regression weights can give equally good predictions for Y. It is in the second step of procedure 3 where a problem might arise. We are now applying the regression equation obtained from step one (using selected cases only) to x and x^2 in the unselected group in order to predict the missing y cases. Thus the range of x scores in the selected sample (utilized in step one) is totally

different from the range of x scores in the unselected sample that is utilized in step two. In step two, then, we are extrapolating to a region (i.e., the unselected Y cases) that is completely outside of the range to which the model was fit (i.e., the selected cases). Therefore, in this case, a large variation in regression coefficients can produce a large variation in the predicted y scores, thereby causing EMSE values for index 3 to be elevated.

Procedure 2, the "reproduced cases" method, took the correct model and substituted z as a proxy for x^2. Assuming that the variable z was not as highly correlated with x as was x^2 in the selected group, the regression weights computed (by regressing y on x and z in the selected group) would have a smaller standard error (than those weights obtained using x and x^2 in a linear regression analysis). While this again would not effect EMSE values obtained using the selected sample, it might indeed effect predicted y values obtained when extrapolating to the unselected group (i.e., missing y cases). Consequently, procedure 2 can be more successful in predicting missing y scores than is procedure 3. This argument is supported by the result that the application of procedure 2 produced smaller EMSE values than that of procedure 3, in the case when the relationship between x and y was linear in form.

This result was also often found when the relationship between x and y was monotonic in form.

It was noted that the performance of index 3 remained constant across all levels of alpha. In other words, the performance of procedure 3 was independent of the relationship between x and y. The explanation for this is that in general, when the x values which define a model are the same as the x values that are utilized in the analysis, the corresponding index will be independent of the relationship between x and y. In this case, since both the underlying model and procedure 3 utilize x and x^2 in predicting missing y cases, the values for index 3 are independent of the relationship between x and y (alpha). This can be demonstrated more specifically as follows:

Given that the underlying relationship between x and y can be described by the quadratic equation demonstrated in equation 1 (where $y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + e$). and procedure 3 is utilized to obtain predicted y scores (i.e., where y is regressed upon x and x^2 to obtain missing y scores), it has been demonstrated that:

$E(\underline{y}_u - H\underline{y}_s)'(\underline{y}_u - H\underline{y}_s)/n_u = E(\underline{y}'A\underline{y})/n_u$

$E(\underline{y}'A\underline{y})/n_u = tr(A\Sigma + \mu'A\mu)/n_u$

where $\mu = x\alpha$; $\Sigma = \sigma^2 I$.

It can be proven that $\mu A \mu /nu = 0$ when the underlying model and the procedure are both based upon the same x scores (in this case x0, x and x^2). Therefore: $E (Y_u - H_s)'(y_u - H_s)/n_u = tr(A)$, and A is not a function of α.

It was hypothesized that method 2, which utilized the x variable would, in some cases, be more successful than method 1 (which uses selected cases only) in predicting missing cases. This result was supported by the data, in the case where the xy relationship was non-monotonic. (The explanation for this will be elaborated on in a following section). However, in the cases where the xy relationships were linear or monotonic, the performance of method 2 was often inferior to that of method 1. One possible explanation for this concerns the manner in which the z variable is defined. The z variable serves as a proxy for x^2. The ideal z variable achieves a balance between being correlated with X (and thus able to predict y), and yet not as highly correlated with x as is x^2. It is not clear that we accomplished the generation of an "ideal z" in the present study. It may be that there are better definitions for z which would yield better estimates of the missing y cases.

There is another possible explanation for the result that the "selected cases" procedure was more

successful in predicting missing y scores than the "reproduced cases" procedure. The introduction of an additional variable, z, requires the estimation of an additional parameter. This, in turn, might produce the larger values obtained for the EMSE of the "reproduced cases" procedure.

Procedure 1, the "selected cases" method, did not utilize the appropriate model to estimate the relationship between x and y, since x was used alone to predict y. However, the performance of this procedure was often superior to that of the other two procedures. While one would expect this method to have the smallest EMSE values in the case when the xy relationship was linear, these results were not expected in the case where the xy relationship was monotonic.Perhaps this can be explained by the fact that method 1 is the simplest of the three methods, and requires the estimation of the smallest number of parameters, thereby resulting in the smallest errors of prediction.

When the relationship between x and y changed from being nonlinear but monotonic to non linear and non-monotonic, procedures 1 and 2 became much less effective in estimating the missing y cases (over all the independent variables). Figure 10 provides an illustration of the two different xy relationships.

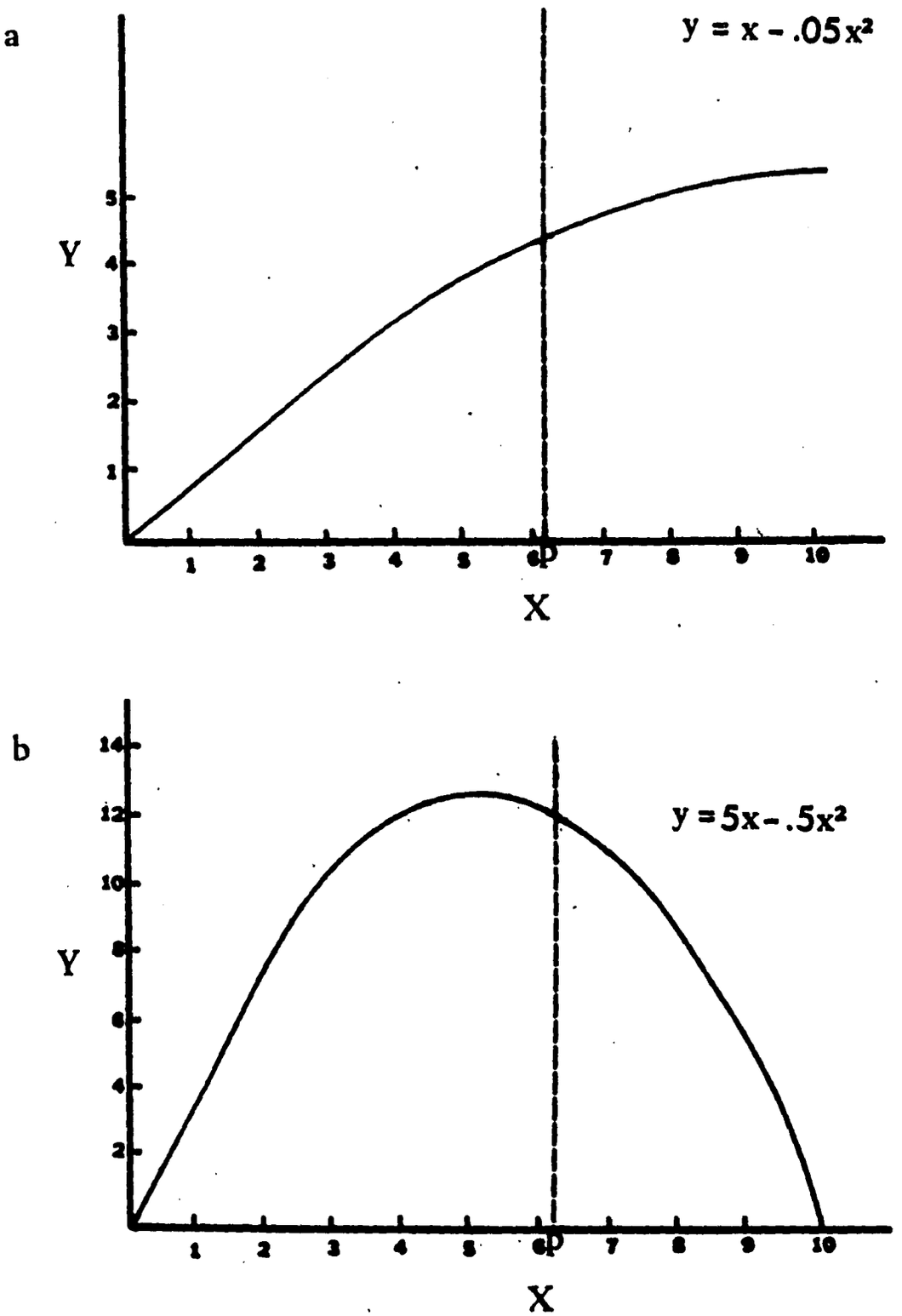


Figure 10. Two sample xy relationships: a. nonlinear but monotonic; b. nonlinear and non-monotonic[a].

[a]p = at a given x value, a proportion, p, is selected.

In (a), the curved line on the graph represents the nonlinear but monotonic relationship between x and y. Those cases to the right of the perpendicular dotted line represent the proportion selected. In (b), the curved line on the graph represents the nonlinear and non-monotonic relationship between x and y. Those cases to the right of the perpendicular dotted line represent the proportion selected.

In both situations depicted by Figure 10 (a) and (b), one is utilizing data to the right of the perpendicular line in an attempt to predict the missing y scores. In both these situations y is regressed on x (or on x and x^2, or on x and z) in the selected group, and the resulting regression weights are used to obtain predicted missing y values. However, in the situation depicted in (b), where the relationship between x and y is parabolic, the estimation of missing y cases can be considerable worse than that of case (a). It is clear than one cannot extrapolate the fitted regression function obtained from the right hand data to the left hand data. This argument is supported by the fact that for procedures 1 and 2, the EMSE values were larger when the relationship between x and y was nonlinear and nonmonotonic, rather than nonlinear but monotonic. However, the EMSE values for procedure 2, when the relationship between x and z was non-monotonic, were

much smaller than the corresponding values obtained when the relationship among x and z was monotonic. Apparently a non-monotonic xz relationship enables one to pick up some cases that are indicative of a parabolic relationship among x and y. Similarly the EMSE values for procedure 3 were (in most cases) much smaller than corresponding EMSE values for procedures 1 and 2, (when the xy relationship was non-monotonic). Thus, the x^2 term might also help to estimate the missing cases that occur on the left side of the parabola depicted in Figure 10.

Certain results relating to the independent variables N, Psel, and Alpha (α) were expected. As the number of cases (N) increased (from n=25 to n=50 to n=100), each of the three procedures became more effective in estimating the missing cases, and the differences in the effectiveness of the three procedures decreased. Similarly, as the percentage of selected cases (Psel) increased, the three procedures became more effective at estimating missing cases. It should be noted that the differences in the index values at each level of n and psel were much more extreme for index 3 (as compared to indices 1 and 2).

An interesting finding in the study concerns the importance of the distribution of x scores. When this was varied, it had a profound effect on the ability of

the three procedures to predict missing y cases. All three procedures were much less successful in predicting missing cases when the distribution of x scores was skewed to the left. Let us consider an explanation for this result. When one is utilizing an x score distribution that is skewed to the left, the majority of cases occur at the upper end of the distribution. All cases at the upper end of the distribution are selected. In this situation, one must deal with a range of scores that is very restricted (see Figure 4). Estimation of missing cases is therefore not very successful. Conversely, when one is considering an x score distribution that is skewed to the right, the majority of cases fall at the lower end of the distribution (see figure 3). Selection of cases at the upper end of the distribution will still result in a wide range of x values for consideration. When a broader spectrum of scores is available, this will result in a more successful estimation for the missing cases.

The performance of indices 1, 2 and 3 can be summarized as follows:

1. In the situation where an underlying relationship among x and y is known to be linear, the utilization of selected x cases alone (method 1) is the method of choice for predicting missing y cases.

2. In a situation where an underlying xy relationship is suspected of being nonlinear, the utilization of an additional variable, (having a nonlinear relationship to y), in conjunction with x, is the method of choice in predicting missing y cases. While it is not clear exactly what this variable should be, the success of this method appears to depend upon the relationship it has to x and y, as well as the underlying relationship between x and y.

3. In a situation where the range of known x values is severely restricted, the performance of all three procedures is unreliable, and the performance of method 3 is especially poor.

Several areas of interest have arisen from this study which can be the subject of future research. First, a further in-depth analysis of the z variable and its properties might prove illuminating. Such a study might compare and contrast z variables having different characteristics (in terms of relationships to y, x, and x^2), in order to determine if there are indeed situations where this type of variable can optimally predict missing cases. It should be possible to find situations where an "optimal z" can provide results (using the "reproduced cases" procedure) that are an improvement over those found using procedure 1 (the "selected cases" procedure).

Another area needing exploration in future research concerns the range of x scores that are utilized to predict missing y scores. As explained previously in this chapter, procedure 3 utilized x and x^2 scores in the selected group to predict y scores in the unselected group. The poor performance of this procedure might be explained by the fact that the x range on which the model was fit (i.e., the selected x cases) was a different range than the one to which we were extrapolating (i.e., the unselected y cases). Therefore, a future study might investigate the use of x and x^2 in the total group to predict the missing y scores. In such a study, procedure 3 might be more effective than it was in this study, since one would be using the x and x^2 scores in both selected and unselected cases to determine the missing y scores(i.e., the y scores in the unselected group).

We will now consider how results from the present study can be applied to educational selection and decision making. Suppose that we are interested in the relationship between scores on an aptitude test, x, and job performance, denoted by y. Let us assume that the job consists of teaching high school courses at private high schools. Suppose that if we had the x and y values for everyone, the relationship between x and y would be parabolic, and could be illustrated by the curve depicted in Figure 7. This kind of curve demonstrates that beyond a certain level of x (aptitude), job performance (y) decreases. Let us consider how, in the situation depicted here, one might find this kind of x-y relationship. Suppose a group of teachers, who are high scorers on an aptitude test (x), lost their jobs, which consisted of teaching college level courese. In need of work, these people enter lower level teaching jobs, where they are assigned to teach lower level courses. They are thus overqualified in their current work situation. The performance of this group of people (who score in the upper range on an aptitude test) will be poorer than that of other teachers having lower scores on the aptitude test due to factors such as boredom, lack of motivation, and dissatisfaction with salary. Hence when looking at the relationship between performance on an aptitude test and job performance, one

might indeed observe a parabolic relationship between x and y, where job performance decreases in the upper range of x.

If we had all the x-y data, we could effectively estimate the x-y relationship utilizing a regression analysis, where y is a function of x and x^2. However, in our example, teachers have been selected for their jobs based upon their scores on the aptitude test, x, and thus we only have performance ratings (y) for those teachers selected for the job. Performance ratings are missing for those applicants not selected for the job. Suppose the school administration is interested in finding out what the job performance for the unselected applicants would be. The data that is observed is demonstrated in Figure 10b, and consists of all those cases to the right of the perpendicular line. We need to utilize these observed performance ratings in order to fill in missing ratings of job performance for the unselected individuals.

We have already stated that we are contemplating a case where the majority of applicants for these teaching positions are overqualified for their jobs. Thus the distribution of x scores (i.e., scores on the aptitude test) will be skewed to the left, with the majority of applicants scoring in the upper range on the aptitude test. In this type of situation, coupled with the fact

that teachers are selected based upon x, the range of x values will be curtailed. We have demonstrated from our study that when the range of x is small, as we have in this case, the results obtained from any analysis will be subject to more error than there would be when performing the same analysis using a wider range of x values. Given a situation where the underlying x-y relationship is nonlinear (and specifically parabolic), and the range of x values is curtailed, we are interested in performing an analysis in order to recover the missing y cases. We could do a simple regression analysis, where we would regress available job performance ratings on aptitude scores and extrapolate back for the unselected population (i.e., procedure 1). However, results from the present study have indicated that a simple regression analysis on selected cases will yield inaccurate results when the x-y relationship is parabolic. If we look at figure 10b, it is apparent that one cannot extrapolate the fitted regression function obtained from the right hand data (by doing a simple linear regression of y on x in the selected group) to the left hand data.

Results from our present study have demonstrated that a regression analysis which utilizes another variable (that has a nonlinear relationship with x), in

conjunction with x to estimate the missing y values will yield more accurate results. For example, often scores on personality inventories are available for use, where the personality traits measured have a nonlinear relationship with performance. (For instance, anxiety has been shown to have a nonlinear. and specifically parabolic, relationship with aptitude). Thus one could regress the job performance measure (y) on aptitude test scores and personality scores for those individuals selected, and use the resulting regression weights to predict the missing job performance values. These estimates would be more accurate than those obtained from a simple linear regression analysis using selected cases.

Appendix A

A FORTRAN COMPUTER PROGRAM TO COMPUTE THE EXPECTED MEAN SQUARED ERROR VALUES FOR PROCEDURES 1, 2, and 3.

FORTRAN PROGRAM

```
      DIMENSION XS(100),XU(100),ZS(100),ZU(100)
      DIMENSION YHATS(100),YHATU(100),ALPHA(3),W(3)
      DIMENSION NN(3),PSEL(3)
      DIMENSION NPROP(10)
      DATA ALPHA/0.0,1.0,-0.05/
      DATA W/0.0,1.0,-0.05/
      DATA NN/25,50,100/
      DATA PSEL/0.6,0.75,0.90/
C
      SIGMA2 = 1.0
C
      WRITE (6,333) ALPHA,W
333   FORMAT(10X, 'ALPHA =' , 3F10.4//10X,
     C'W =',3F10.4///)
      DO 1000 IDIST = 1,3
      READ(5,100) NPROP
100   FORMAT(10I2)
      WRITE (6,101) NPROP
101   FORMAT(1X,'XDIST=',10I3//)
      DO 1 I = 1,3
      DO 1 J = 1,3
      N = NN(I)
      IF(I .EQ. 1 .AND. IDIST .EQ. 1) N = 30
      NS = PSEL(J) * N
      NU = N - NS
      CALL GETDAT(N,NS,NU,ALPHA,W,XS.XU.ZS,ZU,
     CYHATS,YHATU,NPROP,CIDIST)
300   FORMAT(1X,5F10.4)
      XINDEX = XIX(N,NA,NU,ZA,ZU,YHATS,YHATU,SIGMA2)
      ZINDEX = XIZ(N,NS,NU,XS,XU,YHATS,YHATU,ZS,ZU,
     CSIGMA2)
      X2IDEX = XIX2(N,NS,NU,XS,XU,YHATS,YHATU,ZS,ZU,
     CSIGMA2)
      WRITE(6,102) N,NS,NU,XINDEX,ZINDEX,X2IDEX
1     CONTINUE
C
102   FORMAT(3I5,3F10.3)
1000  CONTINUE
      STOP
      END
```

```
      SUBROUTINE GETDAT(N,NS,NU,ALPHA,W,XS,XU,ZS,ZU,
     CYHATS,YHATU,NPROP,IDIST)
C     COMPUTE XS,XU,ZS,ZU,YHATS,YHATU
      DIMENSION ALPHA(3),W(3),XS(100),XU(100),
     CZS(100),ZU(100)
      DIMENSION YHATS(100), YHATU(100)
      DIMENSION DATA(100,3)
      DIMENSION NPROP(10)
      SDZ = 1.0
      SDY = 1.0
      ISTART = 12359
      KOUNT = 0
      DO 1 I = 1,10
      JEND = NPROP(I)
      IF (N .EQ. 50) JEND = JEND/2
      IF (N .EQ. 25 .AND. IDIST .NE. 1) JEND = JEND/4
      IF (N .EQ. 30 .AND. IDIST .EQ. 1) JEND = JEND/4 +1
      DO 1 J = 1,JEND
      KOUNT = KOUNT + 1
      X = I
      XMEANY=ALPHA(1) +ALPHA(2)*X +ALPHA(3)*X**2
      XMEANZ=W(1) +W(2)*X +W(3)*X**2
      CALL GAUSS(ISTART,SDY,XMEANY,DATA(KOUNT,2))
      CALL GAUSS(ISTART,SDZ,XMEANZ,DATA(KOUNT,3))
1     DATA(KOUNT,1)=X
      DO 4 I = 1,NU
      X = DATA(I,1)
      XU(I) = X
      YHATU(I) = ALPHA(1) + ALPHA(2)*X + ALPHA(3)*X*X
4     ZU(I) = DATA (I,3)
      DO 5 I = 1,NS
      X = DATA(NU + I,1)
      XS(I) = X
      YHATS(I) = ALPHA(1) + ALPHA(2)*X + ALPHA(3)*x*x
5     ZS(I) = DATA(NU+I,3)
      RETURN
      END
```

```
      FUNCTION XIX(N,NS,NU,XS,XU,YHATS,YHATU,SIGMA2)
C THIS SUBROUTINE COMPUTES THE EMSE USING X AS THE
C PREDICTOR
      DIMENSION XS(100), XU(100)
      DIMENSION YHATS(100), YHATU(100)
      DIMENSTION H(100,100),HTH(100,100)
      CALL GETHX(NS,NU,XS,XU,H,HTH)
C
      SUM1 = 0.0
      DO 1 I = 1,NU
1     SUM1 = SUM1 + YHATU(I)*YHATU(I)
C
      SUM2 = 0.0
      DO 2 I = 1,NS
      DO 2 J = 1,NS
2     SUM2 = SUM2 + YAHTS(I)*HTH(I,J)*YHATS(J)
C
      SUM3 = 0,0
      DO 3 I = 1,NU
      DO 3 J = 1,NS
3     SUM3 = SUM3 +YHATU(I)*H(I,J)*YHATS(J)
      SUM3 = -2.0*SUM3
C
      SUM = SUM1 + SUM2 + SUM3
C
      TRACE = NU
      DO 4 I=1,NS
 4    TRACE = TRACE + HTH(I,J)
      TRACE + SIGMA2*TRACE
C
      XIX = (TRACE+SUM)/NU
      RETURN
      END
```

```
      FUNCTION XIZ(N,NS,NU,XS,XU,YHATS,YHATU,ZS,ZU,
     CSIGMA2)
C THIS FUNCTION COMPUTES THE EMSE USING X AND Z AS THE
C PREDICTORS
      DIMENSION XS(100),XU(100),ZS(100),ZU(100),
     CYHATS(100),YHATU(100)
      DIMENSION H(100,100), HTH(100,100)
      CALL GETHZNS,NU,XS,XU,ZS,ZU,H,HTH)
C
      SUM1 = 0.0
      DO 1 I = 1,NU
1     SUM1 = SUM1 + YHATU(I)*YHATU(I)
C
      SUM2 = 0.0
      DO 2 I = 1,NS
      DO 2 J = 1,NS
2     SUM2 = SUM2 + YHATS(I)*HTH(I,J)*YHATS(J)
C
      SUM3 = 0.0
      DO 3 I=1,NU
      DO 3 J=1,NS
3     SUM3 =SUM3 + YHATU(I)*H(I,J)*YHATS(J)
      SUM3 = -2.0*SUM3
C
      SUM = SUM1 + SUM2 + SUM3
C
      TRACE = NU
      DO 4 I = 1,NS
4     TRACE = TRACE + HTH(I,I)
      TRACE = SIGMA2*TRACE
C
      XIZ = (TRACE + SUM)/NU
      RETURN
      END
```

```
      FUNCTION XIX2(N,NS,NU,XS,XU,YHATS,YHATU,ZS,ZU,
     CSIGMA2)
C THIS FUNCTION COMPUTES THE EMSE USING X AND X2 AS
C THE PREDICTORS
      DIMENSTION XS(100),XU(100),ZS(100),ZU(100),
     CYHATS(100),YHATU(100)
      DIMENSTION H(100,100), HTH(100,100)
      CALL GETHX2(NS,NU,XS,XU,ZS,ZU,H,HTH)
C
      SUM = 0.0
      DO 1 I = 1,NU
1     SUM = SUM1 + YHATU(I)*YHATU(I)
C
      SUM2 = 0.0
      DO 2 I = 1,NS
      DO 2 J = 1,NS
2     SUM2 = SUM2 + YHATS(I)*HTH(I,J)*YHATS(J)
C
      SUM3 = 0.0
      DO 3 I = 1,NU
      DO 3 J = 1,NS
3     SUM3 = SUM3 + YHATU(I)*H(I,J)*YHATS(J)
      SUM3 = -2.0*SUM3
C
      SUM = SUM1 + SUM2 + SUM3
C
      TRACE = NU
      DO 4 I = 1,NS
4     TRACE = TRACE + HTH(I,J)
      TRACE = SIGMA2*TRACE
C
      XIX2 = (TRACE + SUM)/NU
      RETURN
      END
```

```
      SUBROUTINE GETHX(NS,NU,XS,XU,H,HTH)
C GET H WHEN X IS THE PREDICTOR
      DIMENSION XS(100), XU(100)
      DIMENSION XX2S(100,2),XX2U(100,2),X2T(2,100),
     CS2S(2,2),H(100,100),HTH(100,100),TEMP2(100,2),
     CL(2),M(2)
C
      DO 1 I = 1,NS
      XX2S(I,1)=1.0
      XX2S(I,2) = XS(I)
1     CONTINUE
      DO 2 I = 1,NU
      XX2U(I,1) = 1.0
      XX2U(I,2) = XU(I)
2     CONTINUE
      DO 500 I = 1,2
      DO 500 J = 1,2
      S2S(I,J) = 0.0
      DO 500 K = 1,NS
500   S2S(I,J) = S2S(I,J)+XX2S(K,I)*XX2S(K,J)
      CALL MINV(S2S,2,D,L,M)
C     WRITE(6,100) D,((S2S(I,J),J=1,2),I=1,2)
100   FORMAT('DET = ',F10.3,3(/,1X,3F10.3))
      DO 600 I = 1,NU
      DO 600 J = 1,2
      TEMP2(I,J)=0.0
      DO 600 K = 1,2
      TEMP2(I,J) =TEMP2(I,J)+XX2U(I,K)*S2S(K,J)
600   CONTINUE
      DO 700 I = 1,NU
      DO 700 J = 1,NS
      H(I,J)=0.0
      DO 700 K = 1,2
700   H(I,J)=H(I,J)+TEMP2(I,K)*XX2S(J,K)
      DO 800 I = 1,NS
      DO 800 J = 1,NS
      HTH(I,J)=0.0
      DO 800 K = 1,NU
800   HTH(I,J)=HTH(I,J)+H(K,I)*H(K,J)
      RETURN
      END
```

```
      SUBROUTINE GETHZ(NS,NU,XS,XU,ZS,ZU,H,HTH)
C GET H WHEN X AND Z ARE THE PREDICTORS
      DIMENSION XS(100),XU(100),ZS(100),ZU(100)
      DIMENSION XX2S(100,3),XX2U(100,3)X2T(3,100),
     CS2S(3,3),H(100,100),HTH(100,100),TEMP2(100,3),
     CL(3),M(3)
C
      DO 1 I = 1,NS
      XX2S(I,1)=1.0
      XX2X(I,2)=XS(I)
      XX2S(I,3)=ZS(I)
1     CONTINUE
      DO 2 I = 1,NU
      XX2U(I,1)=1.0
      XX2U(I,2)=XU(I)
      XX2U)I,3)=ZU(I)
2     CONTINUE
C
      DO 500 I = 1,3
      DO 500 J = 1,3
      S2S(I,J)=0.0
      DO 500 K = 1,NS
500   S2S(I,J)=S2S(I,J)+XX2S(K,I)*XX2S(K,J)
      CALL MINV(S2S,3,D,L,M)
C     WRITE(6,100)D,((S2S(I,J),J=1,3),I=1,3)
100   FORMAT( 'DET = ',F10.3,3(/,1X,3F10.3))
      DO 600 I = 1,NU
      DO 600 J = 1,2
      TEMP2(I,J) = 0.0
      DO 600 K = 1,2
      TEMP2(I,J) = TEMP2(I,J) + XX2U(I,K)*S2S(K,J)
600   CONTINUE
      DO 700 I = 1,NU
      DO 700 J = 1,NS
      H(I,J) = 0.0
      DO 700 K = 1,2
700   H(I,J) = H(I,J) + TEMP2(I,K)*XX2S(J,K)
      DO 800 I = 1,NS
      DO 800 J = 1,NS
      HTH(I,J) = 0.0
      DO 800 K = 1,NU
800   HTH(I,J) = HTH(I,J) +  H(K,I)*H(K,J)
      RETURN
      END
```

```
      SUBROUTINE GETHX2(NS,NU,XS,XU,ZS,ZU,H,HTH)
C GET H WHEN X AND X2 ARE THE PREDICTORS
      DIMENSION XS(100),XU(100), ZS(100), ZU(100)
      DIMENSION XX2S(100,3),XX2U(100,3),X2T(3,100),
     CS2S(3.3),CH(100,100),HTH(100,100),TEMP2(100,3),
     CL(3),M(3)
C
      DO 1 I = 1,NS
      XX2S(I,1) = 1.0
      XX2S(I,2) = XS(I)
      XX2S(I,3) = XS(I)*XS(I)
1     CONTINUE
      DO 2 I = 1,NU
      XX2U(I,1) = 1.0
      XX2U(I,2) = XU(I)
      XX2U(I,3) = XU(I)*XU(I)
2     CONTINUE
C
      DO 500 I = 1,3
      DO 500 J = 1,3
      S2S(I,J) = 0.0
      DO 500 K = 1,NS
500   S2S(I,J) = S2S(I,J) + XX2S(K,I)*XX2S(K,J)
      CALL MINV(S2S,3,D,L,M)
C     WRITE(6,100) D, ((S2S(I,J),J=1,3),I=1,3)
100   FORMAT( ' DET = ',F10.3,3(/,1X,3F10.3))
      DO 600 I = 1,NU
      DO 600 J = 1,3
      TEMP2(I,J) = 0.0
      DO 600 K = 1,3
      TEMP2(I,J) = TEMP2(I,J) +XX2U(I,K)*S2S(K,J)
600   CONTINUE
      DO 700 I = 1,NU
      DO 700 J = 1,NS
      H(I,J) = 0.0
      DO 700 K = 1,3
700   H(I,J) = H(I,J) + TEMP2(I,K)*XX2S(J,K)
      DO 800 I = 1,NS
      DO 800 J = 1,NS
      HTH(I,J) = 0.0
      DO 800 K = 1,NU
800   HTH(I,J) = HTH(I,J) + H(K,I)*H(K,J)
      RETURN
      END
//GO.SYSIN DD *
10101010101010101010
322012 8 8 4 4 4 4 4
 4 4 4 4 4 8 8122032
```

References

Afifi, A. A., & Elashoff, R. M. (1966).Missing observations in multivariate statistics: review of the literature. Journal of the American Statistical Association, 61, 595-604.

Beale, E. W., & Little, R. J. A. (1975). Missing values in multivariate analysis. Journal of the Royal Statistical Society, Series B, 37 (1), 129-145.

Breland, H. M. (1978). Population validity and college entrance measures. Research and Development Report (RDE 88-79 No. 2). New York: College Entrance Examination Board.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. Journal of the Royal Statistical Society, Series B, 22, 302-307.

Cohen, A. C., Jr. (1955). Restriction and selection in samples from bivariate normal distributions. Journal of the American Statistical Association, 50, 884-893.

Cohen, A. C., Jr. (1957). Restriction and selection in multinormal distributions. Annals of Mathematical Statistics, 28, 731-741.

Dear, R. E. (1959). A principal component missing data method for multiple regression models (Tech. Rep. No. SP-86). System Development Corporation.

Donner, A. (1982). The relative effectiveness of procedures commonly used in multiple regressions analysis for dealing with missing values. The American Statistician, 36, 378-381.

Frane, J. W. (1976). Some simple procedures for handling missing data in multivariate analysis. Psychometrika, 41 (3), 409-415.

Gleason, T. C. & Staelin, R. (1975). A proposal for handling missing data. Psychometrika, 40 (2), 220-252.

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

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

Gross, A. L. & Bicks, P. (1981). The restriction of range problem. Unpublished manuscript, Department of Education, City University of New York.

Gross, A. L. & Fleischman, L. (1983). Restriction of range corrections when both distribution and selection assumptions are violated. Applied Psychological Measurement, 7 (2), 227-237.

Gross, A. L. & Kagan, E. (1983). Not correcting for restriction of range can be advantageous. Educational and Psychological Measurement, 43 (2), 389-397.

Haitovsky, Y. (1968). Missing data in regression analysis. Journal of the Royal Statistical Society, Series B, 30, 67-82.

Kagan, E. (1977). Estimating correlations from restricted samples: An empirical investigation comparing Bayesian and Classical estimation. Unpublished doctoral dissertation, Department of Education, City University of New York.

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. Royal Society of Edinburgh Proceedings, Section A, 62, 28-30.

Linn, R. L. (1968). Range restriction problems in the use of self selected groups for test validation. Psychological Bulletin, 69, 69-73.

Linn, R. L. (1983). Pearson selection formulas: Implications for studies of predictive bias and estimates of educational effects in selected samples. Journal of Educational Measurement, 20 (1), 1-16.

Linn R. L., Harnisch, D. L., & Dunbar, S. B. (1981). Corrections for range restriction: An empirical investigation of conditions resulting in conservative corrections. Journal of Applied Psychology, 66 (6), 655-663.

Lord, F. M., & Novick, R. M. (1968). Statistical Theories of Mental Test Scores. Reading, Mass: Addison-Wesley.

Novick, M. R., & Jackson, . (1974). Statistical Methods in Educational and Psychological Research. McGraw Hill.

Novick, M. R., & Thayer, D. T. (1969) An investigation of the accuracy of the Pearson selection formulas. (E.T.S. RM-69-22), Princeton, N.J. : E.T.S.

Olson, C. A. & Becker, B. E. (1983). A proposed technique for the treatment of restriction of range in selection validation. Psychological Bulletin, 93 (1), 137-148.

Orchard, T., & Woodbury, M. A. (1972). A missing information principle: theory and applications. Proceedings of the Sixth Berkeley Symposium. Mathematics, Statistics, Probability, 1, 697-715.

Payne, D. A., & McMorris, R. L. (1967). Educational and Psychological Measurement: Contributions to Theory and Practice. Blaisedel Publishing.

Pearson, K. (1903). Mathematical contributions to the theory of evolution. Philosophical Transactions of the Royal Society of London: Series A, 200, 1-66.

Roe, R. A. (1979). The correction for restriction of range and the difference between intended and actual selection. Educational and Psychological Measurement, 39, 551-559.

Rydberg, S. (1963). Bias in Prediction. Stockholm: Almquist and Wiksell.

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

Srinivasen, V., & Weinstein, A. G. (1973). Effects of curtailment on an admissions model for a graduate for a graduate management program. Journal of Applied Psychology, 58, 339-346.

Timm. N. H. (1970). The estimation of variance-covariance and correlation matrices from incomplete data. Psychometrika, 35, 417-438.

Wales, T. J., & Woodland, A. D. (1980). Sample selectivity and the estimation of labor supply functions. International Economic Review, 21, 437-468.

Wilks, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. Annals of Mathematical Statistics, 3, 163-195.

Winkler, R. L., & Hayes, W. L. (1975). Statistics, Probability, Inference and Decision (pp. 117-121). Holt, Rinehart & Winston.

Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. The Empire Journal of Experimental Agriculture, 1, 129-142.