INFORMATION TO USERS

While the most advanced technology has been used to photograph and reproduce this manuscript, the quality of the reproduction is heavily dependent upon the quality of the material submitted. For example:

- Manuscript pages may have indistinct print. In such cases, the best available copy has been filmed.
- Manuscripts may not always be complete. In such cases, a note will indicate that it is not possible to obtain missing pages.
- Copyrighted material may have been removed from the manuscript. In such cases, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, and charts) are photographed by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each oversize page is also filmed as one exposure and is available, for an additional charge, as a standard 35mm slide or as a 17"x 23" black and white photographic print.

Most photographs reproduce acceptably on positive microfilm or microfiche but lack the clarity on xerographic copies made from the microfilm. For an additional charge, 35mm slides of 6"x 9" black and white photographic prints are available for any photographs or illustrations that cannot be reproduced satisfactorily by xerography.

McGanney, Mary Lou

PARAMETER ESTIMATION AND RESTRICTION OF RANGE SELECTIVITY ASSUMPTIONS: MODELLING THE MISSING DATA

City University of New York PH.D. 1987

PARAMETER ESTIMATION AND RESTRICTION OF RANGE SELECTIVITY ASSUMPTIONS: MODELLING THE MISSING DATA

by

MARY LOU MCGANNEY

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.

1987

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.

10/20/86
Date

Chair of Examining Committee

10/20/86
Date

Executive Officer

Dr. Alan L. Gross, Chairman

Dr. David Rindskopf

Dr. Roger Millsap

Supervisory Committee

The City University of New York

Abstract

PARAMETER ESTIMATION AND RESTRICTION OF
RANGE SELECTIVITY ASSUMPTIONS:
MODELLING THE MISSING DATA

by

Mary Lou McGanney

Adviser: Professor Alan L. Gross

In test validation studies, the psychometrician must often work with variables on which there are missing data. In a typical case one has full data on a predictor variable, but only partial data on a criterion variable for the subgroup of selected persons. The missing data can be assumed to be missing 1) at random with respect to all of the variables, or 2) as a function of the full data or predictor variable, or 3) as a function of the missing data variable. The third of these is referred to as the nonignorable case, while the others are referred to as ignorable cases.

It has been shown that one can obtain maximum likelihood estimates of the parameters underlying the predictor and criterion variables in a straightforward way if the data are missing according to certain simple patterns and if the missing data processes are ignorable. Where the missing data patterns are not simple, the likelihood function becomes complex and maximization is not straightforward. Where the data are missing as a function of a missing data variable, the missing data process itself must be modelled and included in the likelihood function. In both cases, an iterative maximization procedure must be used to maximize the likelihood function.

In order to investigate the importance of the missing data assumptions and to explore the use of estimation procedures for the nonstraightforward cases, this work analyzed a single real data set with three variables (with missing data on two) under three different sets of missing data assumptions or models: two ignorable models and one nonignorable. The methodology for obtaining maximum likelihood estimates in the nonstraightforward cases is presented. In addition, for each missing data assumption,

maximum likelihood parameter estimates are reported and a discussion of the problems encountered in using the methodology is included.

As a result of intractable computational difficulties, marginal rather than simultaneous maximum likelihood parameter estimates were obtained in the nonignorable case. Contrary to expectations, the estimates under each set of assumptions were approximately the same, strongly suggesting that the missing data were missing at random.

CONTENTS

LIST OF FIGURES AND TABLES

Figure

Table

CHAPTER I

INTRODUCTION

The psychometrician is often interested in the relationship among two or more variables on which there is incomplete information or missing data. A situation involving missing data occurs when, for example, an educational institution chooses, on the basis of one or more selection variables, to admit a group of individuals from a total applicant group. These variables could include an index of past achievement, a score on an admissions test, performance at an interview and other distinquishing personal information. While usually there is information on each selection variable for all applicants, only the smaller group enters the institution and obtains a grade point average. As a consequence, there are missing data, i.e., no grade point averages, for those

who did not enter. The researcher might be interested in the relationship between an entrance examination such as the Graduate Record Examination on which there are scores for all applicants and first year graduate school performance of those admitted. In general, the researcher is called upon to evaluate a variable as a selection variable, i.e., a variable on the basis of which individuals are chosen or rejected by an institution.

In the simple, though unlikely, situation where all applicants are chosen and all enter a program, there are no missing data. The data in such a case would be as in Figure 1.

X	Z
X	Z
X	Z
.	.
.	.
.	.
X	Z

Figure 1. Two variables with no selection.

where x is the predictor and z the outcome variable. In this case, investigation of the relationship between predictor and outcome variables would be accomplished by calculating the correlations among the variables and/or by calculating the least squares prediction equation. In most admissions situations, however, a selection process operates which results in missing data values. In this more likely case, i.e., where selection occurs, the data would be as in Figure 2.

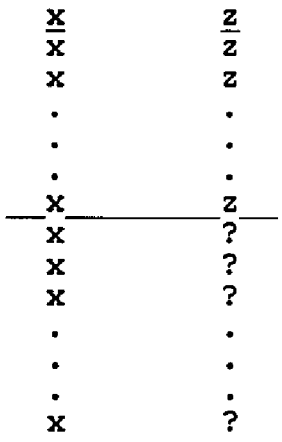

Figure 2. Two variables with selection on one.

If the researcher is interested in the relationships among the variables in the selected group only and does not wish to generalize beyond it, then a product moment correlation or least squares prediction equation using the selected group's data pairs again adequately describes the relationship. More typically, however, the questions of interest concern the usefulness of the predictor variable as a tool in making admissions decisions in the future and the researcher is therefore interested in the relationships among the variables in the population of all possible applicants. In this case the nature of the sampling or selection process must be considered.

Generally, where data are missing, restriction of range is said to occur and as a consequence, a reduction in the estimated variance of the variable for which there are missing data can be expected. If one simply analyzed those cases having both x and z data, and based one's population parameter estimates on the resulting sample statistics, a negatively biased estimate of the correlation coefficient will typically be obtained. Traditionally the approach to correcting for this effect

has been through the use of a formula derived by Pearson (1903) and later studied by Lawley (1943):

$$R_{XZ} = \frac{r_{XZ}}{\left[r_{XZ}^2 + \left(\frac{s_{XS}^2}{s_{Xt}^2}\right) (1.0 - r_{XZ}^2) \right]^{\frac{1}{2}}} \qquad (1)$$

where R_{XZ} is the corrected correlation coefficient between two variables, x and z, where the full range of scores exists for x but not for z; s_{Xt}^2 is a sample estimate of the variance of the unrestricted x population; s_{XS}^2 is an estimate of the variance of the x variable for the selected (restricted) population, i.e., those for whom there are z scores; and r is the correlation between the paired x and z scores where these exist.

Where there are three variables of interest, data may be missing on more than one. If the data are missing as in Figure 3, the restriction of range problem may be addressed through the use of another correction formula (Lord & Novick, 1968):

$$R_{yz} = \frac{r + \left(\frac{s_{xt}^2}{s_{xs}^2} - 1.0\right) r_{xy}\, r_{xz}}{\left\{\left[1.0+\left(\frac{s_{xt}^2}{s_{xs}^2} - 1.0\right) r_{xy}^2\right]\left[1.0+\left(\frac{s_{xt}^2}{s_{xs}^2} - 1.0\right) r_{xz}^2\right]\right\}^{\frac{1}{2}}} \quad (2)$$

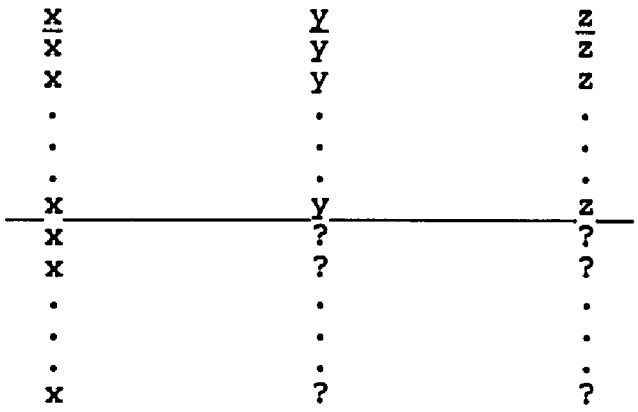

Figure 3. Three variables with a single pattern of selection.

These formulas are based on two underlying assumptions: 1) a distribution assumption and 2) a selection assumption. The first assumes that the regression of the missing data variable(s) on the full data variable is linear and homoscedastic. The second

assumes that the process which results in the missing data is ignorable (Rubin, 1976). In the two variable case (x and z), a selection process is ignorable if, given x, there is no probabilistic relationship between the likelihood of being selected and the outcome variable, z. This occurs when selection is based solely on the x variable, or, if based on other variables too, these are not related to the outcome variable, given knowledge of x. The selection process is non-ignorable when, given x, there is a probabilistic relationship between the likelihood of being selected and the outcome variable, z. It is simple to show that if a non-ignorable selection process is assumed to be ignorable, biased estimates of the strength of the xz relationship will result (Maddala, 1983). In the real world, the ignorable selection process is likely to be the exception.

In the ignorable case where the full and partial data variables are arranged hierarchically as in Figure 4, maximum likelihood estimates of the parameters can be obtained in the following straightforward way:

a) regress the y scores on x,

b) regress the z scores on x and y.

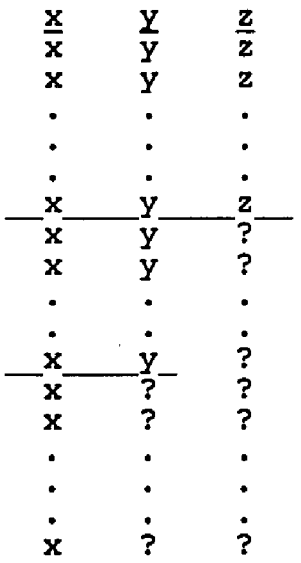

Figure 4. Three variables with two patterns of selection, hierarchically arranged.

For the purpose of discussing and describing the selection process, it is convenient to define a new binary indicator variable, m, standing for whether (m=1) or not (m=0) a subject has a score on the outcome variable. More specifically, the case where m=1 would indicate a subject for whom there is an x score (predictor variable) who is selected and subsequently observed on z (outcome

variable); m=0 would indicate a subject for whom there is an x score, who is not selected and therefore does not have a z score. Using this definition of m, an ignorable selection process can be described as follows:

$$P(m=1 \mid x,z) = P(m=1 \mid x)$$

i.e., given x, there is no probabilistic relationship between the likelihood of being in the program and the z scores. The non-ignorable process can be written:

$$P(m=1 \mid x,z) \neq P(m=1 \mid x)$$

i.e., given x, there is a probabilistic relationship between the likelihood of being in the program and the z scores.

Most relevant work regarding the problem of restriction of range has investigated bias in the correction formula (1) when the distribution and selection process assumptions have been violated. The effect of violating the distribution assumption of linearity and homoscedasticity has been considered by Lord and Novick (1968), Greener and Osburn (1979, 1980), and Novick and Thayer (1969). Their findings indicate that the corrected

correlation estimates using the standard formula are better (less biased) than uncorrected estimates under a variety of simulated violations of these assumptions. Exceptions to this occur when sample sizes are small and when the selection or restriction is extreme.

The selection process and the effect of inappropriately ignoring it has been studied by Forsyth (1971), Gross and Fleishman (1983), Linn (1968), Linn, Harnish and Dunbar (1981) and Rubin (1976). Whereas violations of the distribution assumptions do not seriously affect the estimates obtained using the traditional correction formula, its use when there are violations of the selection process assumption can lead to seriously biased estimates of the size of relationships among variables. Since the traditional procedure using the correction formula for estimating the underlying relationships among the variable where there are missing data gives unbiased estimates only when the selection process is ignorable and since this is rarely the case, it would be desirable to use a method of parameter estimation which is not based on the ignorable selection process assumption.

One can consider maximum likelihood approaches to the problem--approaches which do not require an ignorable selection process and which allows the researcher to take the non-ignorable selection process into account. In order to use these approaches, two different probability models must be introduced.

First, the probability density for each individual's scores on the selection and criterion variables must be formulated. In the case, for example, where there are two selection variables, x and y, and a single outcome or criterion variable, z, the form of this probability density, $P(x_i, y_i, z_i)$ for i=1...N applicants, must be provided. Second, the selection process (or processes) must be modelled. This takes the form of a conditional probability distribution describing the probability that individuals were selected given the data: $P(m=1 \mid x,y,z)$. The researcher could argue that the missing data are missing as a monotonic function of the variables of interest, i.e., the higher the scores, the more likely they are to be present. Or, instead, it could be argued that a non-monotonic relationship exists as would be the case where scores from the low and high ends of the

distribution are less likely to be seen than those from the middle range. Whatever the case, a mathematical statement describing the conditional probability of selection given the data must be provided. With these models of how the variables are related to one another and with the data, the maximum likelihood estimation procedures then provide estimates of the underlying parameters.

Work in applying the maximum likelihood approaches has been done in econometrics (see Maddala,1983) but has not been widely applied to psychometrics. Also remaining to be investigated are actual applications of the maximum likelihood techniques to real and complex data sets. The problems which are likely to be encountered when applying the methods to real data sets can be viewed as of two kinds. In the first place, real situations can be expected to pose modelling difficulties. A real situation might be expected to involve more than one predictor variable, raising the possibility that data are missing on more than one variable. As a consequence, there are more probabilistic relationships which must be modelled and employed, and the manageability of such techniques must be

explored. In the same vein, the use of simplifying assumptions in formulating the problem may be recommended by the researcher and the use and justification of such assumptions must be investigated.

The second kind of problem concerns the actual functioning of the maximum likelihood estimation procedure. To begin with, where the function to be evaluated is a non-linear function of the parameters, one must typically employ an iterative method of estimation, and therefore the appropriateness of the chosen starting values must be assured. Furthermore, the researcher must consider whether or not special difficulties are posed for maximization by complicated functions with many parameters to be estimated.

It is the purpose of this paper to address these issues through the analysis of a real data set consisting of scores on three psychometric variables: a graduate school admissions examination (GRE), undergraduate grade point average (UGPA), and graduate school grade point average (GGPA), for those students who applied for admission to graduate study at a unit of the City University of New York. All subjects have UGPA. Since the

GRE is not required, only some applicants submit these scores and therefore there will be missing data on this variable as well as on the GGPA. Thus, two selection processes will be taking place and two binary indicator variables, m_y and m_z, will be created.

The data set can be represented as in Figure 5:

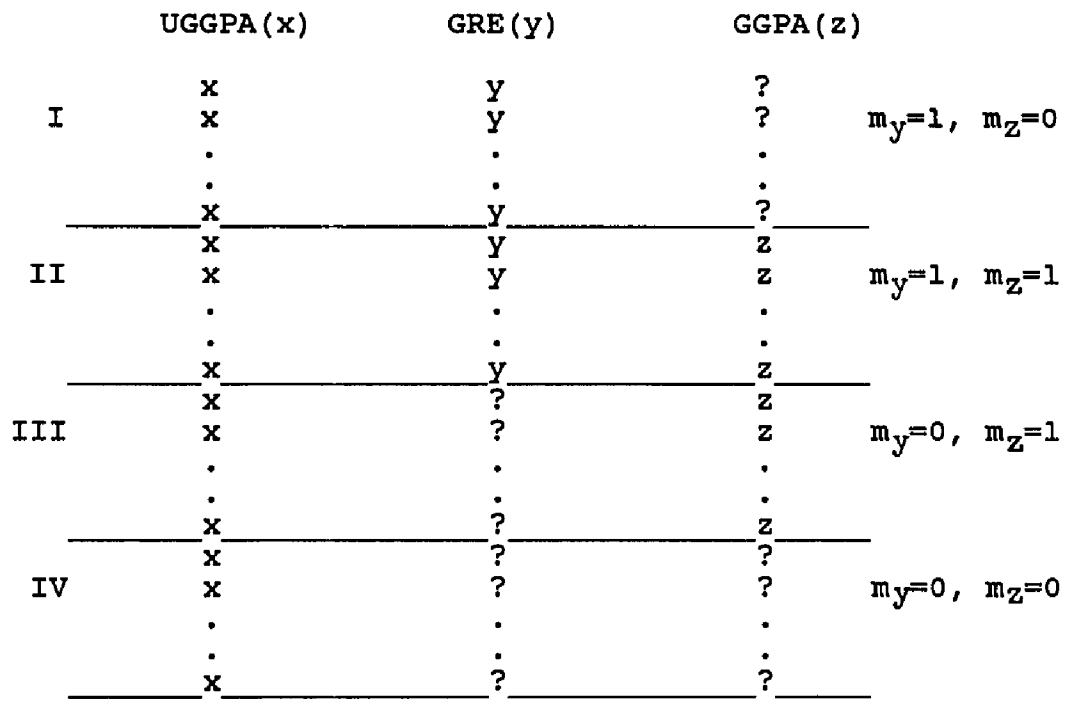

Figure 5. Representation of data set.

where m_y=1 if an individual submits GRE scores (y), m_y=0 if not; m_Z=1 if the individual enters the program and therefore has a graduate grade point average; m_Z=0 if not. The question marks are located where data are missing. The Roman numerals refer to groups distinguishable by their patterns of data and missing data.

Several questions would be asked by the psychometrician concerning these data: 1) how well does undergraduate grade point average, x, predict graduate school performance, z? 2) how well does performance on the Graduate Record Exam, y, predict z? and 3) is prediction of z improved by using both predictors, x and y, together? Answers to these questions require estimation of nine underlying parameters, i.e., μ_x, σ_x^2, $\beta_{0y|x}$, $\beta_{1y|x}$, $\sigma^2_{y|x}$, $\beta_{0z|x}$, $\beta_{1z|x}$, $\sigma^2_{z|x}$, $\sigma_{yz|x}$, which easily can be transformed to get the following set of parameters: μ_x, σ_x^2, σ_y^2, μ_y, μ_z, σ_z^2, σ_{xy}, σ_{xz}, σ_{yz}.

Generally stated, our objective will be to apply the maximum likelihood approach to estimating the underlying parameters of this data set with its patterns of missing

values. For this purpose, we will develop the likelihood functions for our data in the most general case and then modify the functions to conform to three sets of assumptions concerning the nature of the selection processes. Using the functions adapted for our assumptions, we will do three analyses of the data. First we will assume that there are two ignorable selection processes and that the missing GRE scores and graduate GPAs are missing independent of the undergraduate GPA. Second, we will assume that there are two ignorable selection processes but that the missing data are missing as a function of undergraduate GPA. Third, we will assume that in addition to both selection processes being dependent on undergraduate GPA, one of the selection processes is non-ignorable. More specifically, we will argue and assume the following: (a) that the selection process involving GRE scores (m_y) is ignorable, (b) that the selection process involving graduate GPA (m_z) is non-ignorable, and (c) that the best description of this latter selection process is that it is an increasing monotonic function of x and z.

CHAPTER 2

LITERATURE REVIEW

The general problem which we are investigating is one of estimating the relationships among a set of variables when complete data are not observed. Such a case typically arises, for example, in test validations studies where often there are complete data on a predictor variable, x, but incomplete data on a criterion variable, y, due to loss of subjects (Lord & Novick, 1968). The process through which subjects are lost is known as the selection process.

The data for the case where there is a single predictor variable, x, and a single criterion variable, y, (case A) are represented schematically in Figure 6.

x	y	m
x	y	1
x	y	1
.	.	.
.	.	.
.	.	.
x	y	1
x	?	0
x	?	0
.	.	.
.	.	.
.	.	.
x	?	0

Figure 6. Case A.

where x is the predictor variable, y the criterion variable, and m is a binary indicator variable representing the selection process and standing for whether (m=1) or not (m=0) a subject has a score on the criterion variable, y. This case arises when, for example, a graduate school department wishes to evaluate the Graduate Record Examination (GRE) as a predictor of graduate school achievement. There are GRE scores for all applicants, N, but achievement scores only for those (n_S) accepted and enrolled students; typically $n_S < N$.

Another pattern of data (case B) arises when there

is a single predictor variable, x, and two criterion variables, y and z, both of which have the same pattern of missing data. Case B is represented in Figure 7. In case

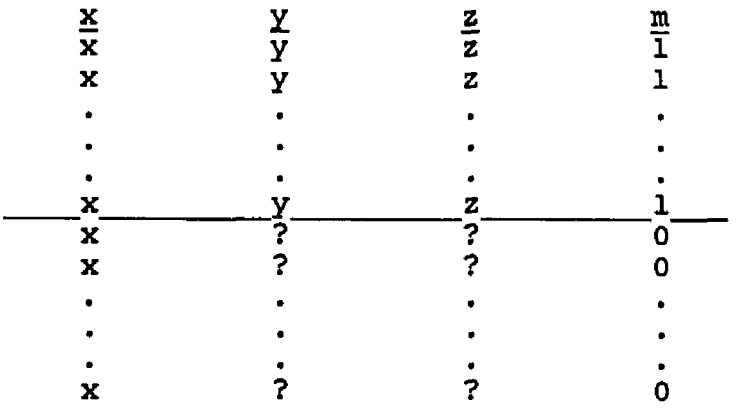

Figure 7. Case B.

B, a single selection process is responsible for the missing data on both y and z. Such a case would arise, for example, when a school wishes to evaluate the GRE as a predictor and has two criteria of interest, e.g., overall grade point average (GPA) and GPA for area of concentration.

While the data patterns described in cases A and B have typically been discussed in the psychometric literature (see for example, Lord & Novick, 1968), more complex patterns frequently emerge from real life situations. The data for the variables in our study are schematically represented by Figure 8.

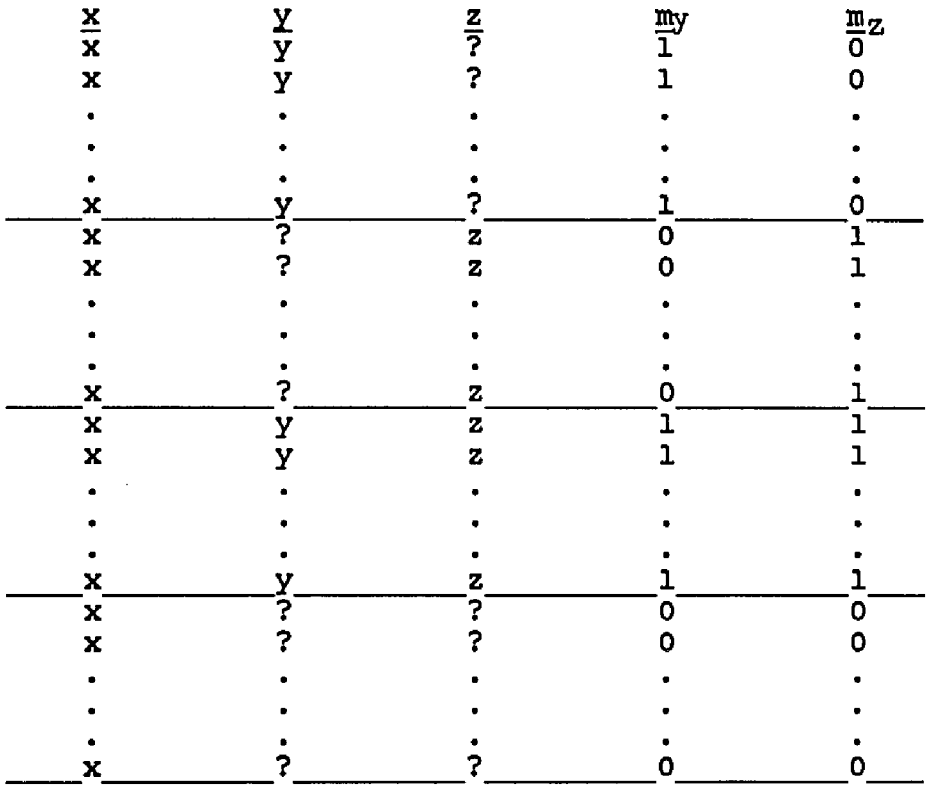

Figure 8. Schematic representation of study's data set.

There are two predictor variables, x and y, and a single criterion variable, z; in addition, there are two selection process indicator variables, m_y and m_z . The m_y variable refers to the selection process resulting in data missing on the y variable and the m_z variable refers to a second selection process resulting in data missing on the z variable. The problem is to estimate the parameters relating the x, y, and z variables.

One general approach to parameter estimation where there are missing data uses the method of maximum likelihood estimation. The maximum likelihood parameter estimates are defined to be those which are most probable, given the data, and, in large samples, have several desirable statistical properties (Hogg & Craig,1970).

In order to obtain the maximum likelihood estimates of the underlying parameters, one must specify the joint likelihood of the sample data. The likelihood, L, of the data consists of the joint probability density for the observed scores for each of N subjects. In general, we can denote the scores for subject i on p variables as v_{1i} ,

v_{2i} ,...,v_{pi} . In addition, we can denote the presence or absence of a score on a variable, v_p , by a score of 1 or 0 on an indicator variable, m_p .

$$L=\prod_{i=1}^{N} p(v_{1i},v_{2i},\ldots,v_{pi},m_{1i},m_{2i},\ldots,m_{pi}\,|\,\underline{\theta},\underline{\phi}) \qquad (3)$$

A subject who has a m_p score of 1 will have a score on v_p. Where m_p equals 0, however, there will be a corresponding gap in the likelihood function due to data being missing on v_p . The parameters underlying the distribution of the variables, $\underline{v}_1,\underline{v}_2,\ldots,\underline{v}_p$, are denoted by $\underline{\theta}$. The parameters underlying the selection process, i.e., the conditional distribution of the m variable(s) given the v variables, are denoted by $\underline{\phi}$. The likelihood function is a function of both $\underline{\theta}$ and $\underline{\phi}$, i.e., $L=f(\underline{\theta},\underline{\phi})$, and is therefore mathematically maximized with regard to the parameters in $\underline{\theta}$ and $\underline{\phi}$. Following the practice of Rubin (1976), we assume the parameters in $\underline{\theta}$ and $\underline{\phi}$ to be distinct.

It is convenient and equivalent to work with the logarithm of the likelihood function:

$$\ln L = \sum_{i=1}^{N} \ln p(v_{1i}, v_{2i}, \ldots, v_{pi}, m_{1i}, m_{2i}, \ldots, m_{pi} \mid \underline{\theta}, \underline{\phi}) \quad (4)$$

Of further convenience is a separation of the log likelihood function into two parts: 1) the probability distribution of the v_{ij} variables, $i=1,2,\ldots,N$, $j=1,2,\ldots,p$, and 2) the probability model(s) for the selection processes, i.e., the probability models for the m variables, given the v variables. The likelihood function (3) then is as follows:

$$L = \prod_{i=1}^{N} [p(v_{1i}, v_{2i}, \ldots, v_{pi} \mid \underline{\theta})] \; X \; [p(m_{1i}, m_{2i}, \ldots, m_{pi} \mid \underline{\theta}, \underline{\phi}, v_{1i}, v_{2i}, \ldots, v_{pi})] \quad (5)$$

and the log likelihood function (4) becomes:

$$\ln L = \sum_{i=1}^{N} \ln p(v_{1i}, v_{2i}, \ldots, v_{pi} \mid \underline{\theta}) + \sum_{i=1}^{N} \ln p(m_{1i}, m_{2i}, \ldots, m_{pi} \mid \underline{\theta}, \underline{\phi}, v_{1i} \; v_{2i} \ldots v_{pi}) \quad (6)$$

In general, the estimates of the $\underline{\theta}$ parameters

depend 1) upon the assumptions concerning the distribution forms of the v variables, and 2) upon the probability models for the selection process of the m variables. The nature of the assumptions can be explored for a variety of missing data situations.

Case 1 Ignorable selection process: $p(m=1|x,y)=p(m=1)$

Suppose the case where there are paired x y data for n_1 subjects and only x data for n_2 subjects. This situation is portrayed in Figure 6. For this case, let it be assumed that the missing y data are missing completely at random: $p(m=1|x,y) = p(m=1)$. Not only is the selection process ignorable since the process is independent of the criterion variable, y, but it is also independent of the predictor variable, x. Further suppose that the x and y scores are from a bivariate normal distribution. This case might arise if a school wished to compare two achievement tests, one given to all students and the other given only to a random sample of students from the school.

Let us separate the likelihood function into two components, L_1 and L_2 , where L_1 gives the likelihood function for group 1 (n_1 subjects having both x and y

scores) and where L_2 gives the likelihood for group 2 (n_2 subjects having x scores only) are as follows:

$$L_1 = \prod_{i=1}^{n_1} p(x_i|\underline{\Theta})\,p(y_i|\underline{\Theta}, x_i)\,p(m_i=1|\underline{\Theta}, x_i, y_i)$$

$$= \prod_{i=1}^{n_1} p(x_i|\underline{\Theta})\,p(y_i|\underline{\Theta}, x_i)\,p(m_i=1|\underline{\Phi}) \qquad (7)$$

$$L_2 = \prod_{i=1}^{n_2} p(x_i|\underline{\Theta})\,p(m_i=0|\underline{\Phi}, x_i)$$

$$= \prod_{i=1}^{n_2} p(x_i|\underline{\Theta})\,p(m_i=0|\underline{\Phi}) \qquad (8)$$

The logarithm of the likelihood function, ln L, is then ln L = ln L_1 + ln L_2 , where

$$\ln L_1 = \sum_{i=1}^{n_1} \ln p(x_i|\underline{\Theta})\,p(y_i|x_i, \underline{\Theta}) + \sum_{i=1}^{n_1} \ln p(m_i=1|\underline{\Phi}) \qquad (9)$$

$$\ln L_2 = \sum_{i=1}^{n_2} \ln p(x_i|\underline{\Theta}) + \sum_{i=1}^{n_2} \ln p(m_i=0|\underline{\Phi}). \qquad (10)$$

It should be noted that $\underline{\Theta}$ and $\underline{\Phi}$ appear in separate terms. In this case it is possible to express the

likelihood function as the sum of two functions: $L = f(\underline{\Theta}, \underline{\phi}) = f(\underline{\Theta}) + f(\underline{\phi})$ since the parameters underlying the x y distribution, i.e., those in $\underline{\Theta}$, and the parameters underlying the selection process, i.e., those in $\underline{\phi}$, appear as separate factors. Thus, given that $\underline{\Theta}$ and $\underline{\phi}$ are distinct parameters, there is no need to jointly maximize $f(\underline{\Theta}\ \underline{\phi})$. In addition, since n_1 can be considered to be a random sample from the population of interest, one could ignore group 2 and estimate $\underline{\Theta}$ by maximizing the likelihood function for group 1 given the group's x and y data; i.e., one maximizes

$$L(\Theta) = \sum_{i=1}^{n_1} \ln p(x_i | \underline{\Theta}) p(y_i | x_i, \underline{\Theta}) \qquad (11)$$

Assuming x and y to be bivariate normal, the maximum likelihood estimates would be as follows:

$\hat{\mu}_x$	$= \bar{x}_1$	=mean of the n_1 x scores
$\hat{\sigma}_x^2$	$= s_{x1}^2$	=variance of the n_1 x scores
$\hat{\beta}_{oy}$	$= b_{0y}$	=least squares regression constant for predicting y from x, computed from the x y data of the n_1 subjects

(12)

$\hat{\beta}_{1y} = b_{1y}$ =least squares regression coefficient for predicting y from x, computed from the x y data of the n_1 subjects

$\hat{\sigma}^2_{y|x} = s^2_e$ =the residual variance computed from the x y data of the n_1 subjects

The population correlation, ρ_{xy}, is a simple function of the parameters contained in $\underline{\theta}$:

$$\rho_{xy} = \frac{\beta_{1y} (\sigma_x^2)^{\frac{1}{2}}}{(\beta_{1y}^2 \sigma_x^2 + \sigma_{y|x}^2)^{\frac{1}{2}}} \qquad (13)$$

The maximum likelihood estimate of ρ_{xy} is simply obtained by replacing the parameters in (13) by their maximum likelihood estimates:

$$\hat{\rho}_{xy} = \frac{b_{1y} s_x}{(b_{1y}^2 s_x^2 + s_e^2)^{\frac{1}{2}}} \qquad (14)$$

$$= r_{xy}$$

where r_{xy} is the product moment correlation between x and y computed in the n_1 subjects.

Of course it would be better to use all of the information, i.e., the x scores in group 2 as well as the x and y scores in group 1, and therefore preferable to maximize the likelihood function:

$$\ln L(\underline{\theta}) = \sum_{i=1}^{n_1} \ln p(x_i | \underline{\theta}) p(y_i | x_i, \underline{\theta}) + \sum_{i=1}^{n_2} \ln p(x_i | \underline{\theta}) \tag{15}$$

$$= \sum_{i=1}^{n_1+n_2} \ln p(x_i | \underline{\theta}) + \sum_{i=1}^{n_1} \ln p(y_i | x_i, \underline{\theta})$$

Again, it should be noted that $\underline{\theta}$ and $\underline{\phi}$ are in separate additive factors and that, for the purpose of estimating $\underline{\theta}$, $\underline{\phi}$ can be ignored. The maximum likelihood estimates of the $\underline{\theta}$ parameters would be as in (12) except

$\hat{\mu}_x = \bar{x}_t$ =mean of the $n_1 + n_2$ scores

$\hat{\sigma}_x^2 = s_{xt}^2$ =variance of the $n_1 + n_2$ scores.

The population correlation would be again as in (13). In this case, however, replacing the parameters in (13) by their maximum likelihood estimates gives the following:

$$\rho_{xy} = \frac{b_{1y} s_{xt}}{(b_{1y}^2 s_{xt}^2 + s_e^2)^{\frac{1}{2}}} \qquad (16)$$

$$= \frac{r_{xy}}{[r_{xy}^2 + (\frac{s_{x1}^2}{s_{xt}^2})(1.0 - r_{xy}^2)]^{\frac{1}{2}}}$$

where r_{xy} is the product moment correlation between x and y computed for group 1; s_{x1}^2 is the variance of x computed for group 1, and s_{xt}^2 is the variance of x computed for all x subjects. In other words, estimates of the x y relationship are made using the x y data from group 1 and estimates of the parameters of the x distribution are made using the x scores of all subjects.

The second equality in (16) is easily recognizable as the widely used correction formula (1) for estimating the x y correlation when data are missing on y (Lord & Novick, 1968). The maximum likelihood estimate for ρ_{xy} given by (16) was originally derived by Cohen (1955). It should be noted that in the case described here where data are missing completely at random, i.e., independently of both x and y, both s^2_{x1} and s^2_{xt} are consistent estimates of σ^2_x. Thus in cases where the sample sizes n_1 and $n_1 + n_2$ are not "large", the ratio s^2_{x1} / s^2_{xt} will typically be close to 1.0. In this case, the estimates for given by (14) and (16) will be similar.

Case 2 Ignorable selection process: $p(m=1|x,y)=p(m=1|x)$

Now suppose the case where again, from a bivariate normal distribution, there are paired xy data for n_1 subjects, and only x data for n_2 subjects (as in Figure 6). Assume here, however, that the missing y data are missing as a function of x: $p(m=1|x,y) = p(m=1|x) \neq p(m=1)$. Such a situation occurs, for example, when a cut off score on a qualifying examination, x, is used to

determine selection and there are no other variables influencing the process. The selection process is independent of y but not of x. Thus the selection process is ignorable as defined earlier. Generally stated, selection has been made only on the basis of a score on the x variable, or, if there are other variables on the basis of which the selection is made, they are not probabilistically related to y. The log likelihoods for group 1 having n_1 subjects and for group 2 having n_2 subjects are as follows:

$$\ln L_1 = \sum_{i=1}^{n_1} \ln p(x_i|\underline{\Theta})p(y_i|x_i,\underline{\Theta}) + \sum_{i=1}^{n_1} \ln p(m_i=1|x_i,\underline{\Phi})$$

$$\ln L_2 = \sum_{i=1}^{n_2} \ln p(x_i|\underline{\Theta}) + \sum_{i=1}^{n_2} \ln p(m_i=0|x_i,\underline{\Phi}) \qquad (17)$$

Again, as in case 1, $\underline{\Theta}$ and $\underline{\Phi}$ appear in separate terms and it is possible to express the likelihood function as the sum of two functions: $L(\underline{\Theta}, \underline{\Phi}) = f_1(\underline{\Theta}) + f_2(\underline{\Phi})$. Given that $\underline{\Theta}$ and $\underline{\Phi}$ are distinct parameters,

there is no need to jointly maximize $f(\underline{\Theta}, \underline{\Phi})$.

The maximum likelihood estimates of the parameters would be as follows:

$\hat{\mu}_x = \bar{x}_t$ =mean of the $n_1 + n_2$ x scores

$\hat{\sigma}_x^2 = s_{xt}^2$ =variance of the $n_1 + n_2$ x scores

(18)

$\hat{\beta}_{oy} = b_{0y}$ =least squares regression constant for predicting y from x, computed from the x y data of the n_1 subjects

$\hat{\beta}_{1y} = b_{1y}$ =least squares regression coefficient for predicting y from x, computed from the x y data of the n_1 subjects

$\hat{\sigma}_{y|x}^2 = s_e^2$ =the residual variance computed from the x y data of the n_1 subjects.

The maximum likelihood estimate of ρ_{xy} is identical to (16). Again, this would be found as follows. Since, by definition, the population parameter is:

$$\rho_{xy} = \frac{\beta_{1y}\,(\sigma_x^2)^{1/2}}{(\beta_{1y}\sigma_x^2 + \sigma_{y|x}^2)^{1/2}}$$

then an estimate of the parameter is:

$$\rho_{xy} = \frac{b_{1y}s_{xt}}{(b_{1y}^2 s_{xt}^2 + s_e^2)^{\frac{1}{2}}}$$

$$= \frac{r_{xy}}{[r_{xy}^2 + (\frac{s_{x1}^2}{s_{xt}^2})(1.0-r_{xy}^2)]^{\frac{1}{2}}}$$

In the earlier case where $p(m|x) = p(m)$, consistent estimates of the population parameters $\hat{\mu}_x$ and $\hat{\sigma}_x^2$ could be based on group 1 only or on the total N x scores. In this case, where the y scores are missing as a function of the x scores, the selected group will not provide a random sample of the x scores. If one bases one's estimates of $\hat{\mu}_x$

and $\hat{\sigma}_x^2$ on the selected group only, the estimates will be biased as will the population correlation, $\hat{\rho}_{xy}$. In order to obtain unbiased estimates of $\hat{\mu}_x$ and $\hat{\sigma}_x^2$, one must use x data from the total N. The other estimates would be based on the data from group 1.

It is this case (case 2) which psychometricians have most frequently examined when investigating the restriction of range problem. The correction formula (1) was derived to correct for restriction of range when there are full data on one variable and partial data on a second (Pearson, 1903). This formula, and the adaptation of it (2), are based on two underlying assumptions: 1) that the regression of y on x is linear and homoscedastic (distribution assumption) and 2) that the process which results in the missing data is ignorable. The robustness of the formulas has been widely investigated when these assumptions are violated.

Studies in which violations of the distribution assumptions have been examined indicate that the estimates obtained using the correction formula are less biased than uncorrected estimates (Greener & Osburn, 1979, 1980; Gross & Kagen, 1983; Novick & Thayer, 1969). Exceptions to this

finding occur when sample sizes are small or selection is extreme i.e., more than 20% of the observations are missing. Studies involving violations of the selection process assumption suggest that the correction formula will typically provide estimates that are seriously biased (Gross & Fleishman, 1983; Linn, 1968; Linn, Harnish & Dunbar, 1981).

Case 3 Non-ignorable selection process: $p(m=1|\underline{x},y) \neq p(m=1|\underline{x})$

Now let us suppose that the selection process is non-ignorable, i.e., that, given x, there remains a probabilistic relationship between being in the selected group (m=1) and the y variable. This occurs, for example, when selection into the group has been made on the basis of several predictor variables, x_1, $x_2,\ldots,x_p$ but only x_1 is observed. The other or additional variables x_2, x_3, $,\ldots,x_p$ can be referred to as $\underline{x}_a^*$. If, given x_1, there is no probabilistic relationship between y and $\underline{x}_a^*$, the selection process is ignorable. However, if given x, y is related to $\underline{x}_a^*$, then the selection process is not

ignorable.

In the ignorable case, the maximum likelihood estimates for the parameters underlying the x_1 y distribution are given in (12) or (19). However, when there is a probabilistic relationship between y and $\underline{x}_a^*$ given x_1, the regression weights, b_{0y} and b_{1y} will be biased estimates of $\hat{\beta}_{0y}$ and $\hat{\beta}_{1y}$ and will lead to biased estimates of ρ_{xy} (Madalla, 1983). More specifically, where $p(m \mid x, y) = p(m \mid x)$,

$$E(y|x,m=1) = \beta_0 + \beta_1 x ; \qquad (19)$$

but when $p(m \mid x\ y) \neq p(m \mid x)$,

$$E(y|x,m=1) = \beta_0 + \beta_1 x_1 + \beta_2 f(x) \qquad (20)$$

where f(x) is a function of x_1. In the econometric literature, the error of ignoring x when estimating the relationship between x and y is referred to as specification error (see Heckman, 1979).

In the non-ignorable case, the likelihoods for group 1 with n_1 subjects having both x and y scores and for group 2 with n_2 subjects having x scores only are as follows:

$$L = \prod_{i=1}^{n_1} p(x_i|\underline{\theta})p(y_i|x_i,\underline{\theta})p(m_i=1|x_i,y_i,\underline{\phi})$$

$$L = \prod_{i=1}^{n_2} p(x_i|\underline{\theta})\int_{y_i} p(m_i=0|x_i,y_i,\underline{\phi})p(y_i|x_i,\underline{\theta}) \qquad (21)$$

The log likelihood functions for each group are:

$$\ln L_1 = \sum_{i=1}^{n_1} \ln p(x_i|\underline{\theta}) + \sum_{i=1}^{n_1} \ln p(y_i|x_i,\underline{\theta}) + \sum_{i=1}^{n_1} \ln p(m_1=1|x_i,y_i,\underline{\phi})$$

$$\ln L_2 = \sum_{i=1}^{n_2} \ln p(x_i|\underline{\theta}) + \sum_{i=1}^{n_2} \ln p(m_i=0|x_i\,\underline{\phi}\,\underline{\theta}) \qquad (22)$$

The complete log likelihood function is:

$$\ln L = \sum_{i=1}^{n_1+n_2} \ln p(x_i|\underline{\Theta}) + \sum_{i=1}^{n_1} \ln p(y_i|x_i,\underline{\Theta})p(m_i=1|x_i,y_i,\underline{\Phi}) + \sum_{i=1}^{n_2} \ln p(m_i=0|x_i,\underline{\Theta},\underline{\Phi}) \quad (23)$$

It should be noted that in this case where it is assumed that the selection process is not independent of the y, variable, $\underline{\Theta}$ and $\underline{\Phi}$ appear in the same term and cannot be treated as separate factors. In order to estimate the parameters underlying the x y relationship, it therefore is necessary to jointly maximize $\underline{\Theta}$ and $\underline{\Phi}$. When the selection process was ignorable and $\underline{\Theta}$ and $\underline{\Phi}$ were in separate additive factors, the maximum likelihood estimates for $\underline{\Theta}$ could be obtained by maximizing (11). In this case, where $\underline{\Theta}$ and $\underline{\Phi}$ are not separable, it is necessary to provide a mathematical description of the x, y, and m joint distribution: $p(x, y, m|\underline{\Theta}\ \underline{\Phi})$ or equivalently, $p(x|\underline{\Theta})p(y|x\ \underline{\Theta})p(m|x,y,\underline{\Phi}\ )$. This requires

the addition of a model for $p(m \mid x,y, \underline{\phi})$, i.e., the selection process.

Madalla (1983), in reviewing work in econometrics, discusses different non-ignorable selection situations and the models which have been proposed to mathematically describe the selection processes. One non-ignorable selection process occurs when x scores are available for a sample, N, but y scores are available only if the y score is above (or only if it is below) a certain y score, the cut off score or c, where c is a known constant. This is shown in Figure 9. Such a situation could occur in education, for example, where detailed family income information is often available only if family income is below a certain level, but scholastic aptitude information is available for all students. The psychometrician may wish to consider the relationship between aptitude and family income in the general population.

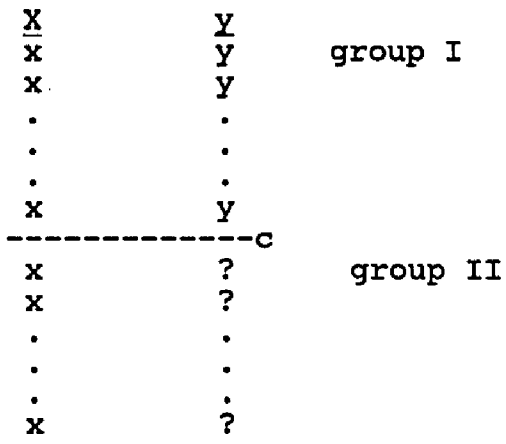

Figure 9. Missing data where there is a cutoff point y_c.

It can be seen in Figure 10 that the x y data for group one is systematically different from the x y data set that would have been observed for all subjects if there were no missing data. It is further evident that, as

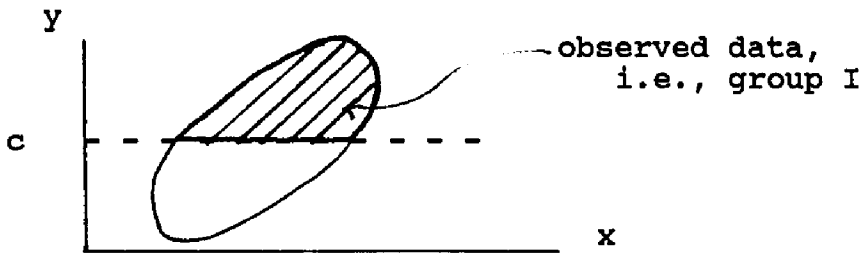

Figure 10. Distortion of x y data plot as a result of non-ignorable selection.

a consequence, the true regression line for the population and the estimated regression line calculated on the basis of the observed x y pairs can be significantly different. This contrast is illustrated in Figure 11.

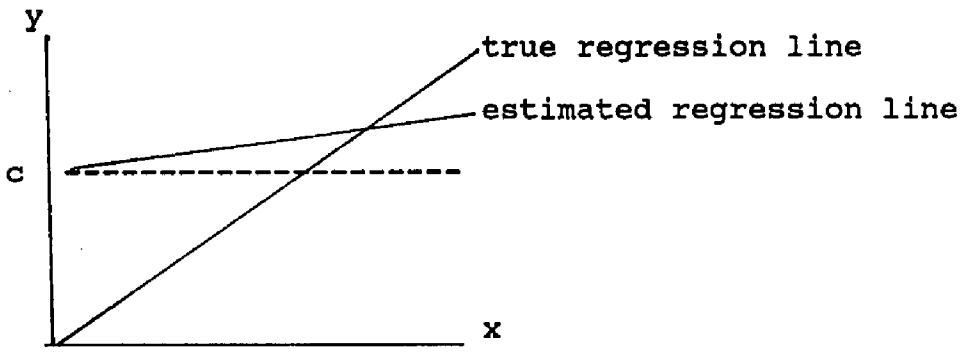

Figure 11. Illustration of likely differences in regression line in the case of non-ignorable selection.

The model for this non-ignorable selection process is as follows:

$$p(m=1|x,y) = 1 \quad \text{if } y \geq c$$

$$= 0 \text{ otherwise}$$

The Tobit model (Tobin, 1958) describes this process in the special case when c is equal to zero:

$$p(m=1|x,y) = p(y \geq 0|x,y).$$

The likelihood functions for group I with paired x,

y data and m=1, and for group II with x data only and m=0, are as follows:

$$\ln L_1 = \sum_{i=1}^{n_1} \ln p(x_i|\underline{\Theta})p(y_i|x_i\underline{\Theta})p(m_i=1|x_i,y_i,\underline{\Phi}); \quad (24)$$

since for group I, $p(m_i=1|x_i,y_i,\underline{\Phi})=1$,

$$= \sum_{i=1}^{n_1} \ln p(x_i|\underline{\Theta})p(y_i|x_i\underline{\Theta}).$$

$$\ln L_2 = \sum_{i=1}^{n_2} \ln \int_y p(x_i|\underline{\Theta})p(y_i|x_i\underline{\Theta})p(m_i=0|x_i,y_i,\underline{\Phi}) \, dy$$

$$= \sum_{i=1}^{n_2} \ln p(x_i|\underline{\Theta}) + \quad (25)$$

$$\sum_{i=1}^{n_2} \ln \int_{-\infty}^{\infty} p(y_i|x_i,\underline{\Theta})p(m_i=0|x_i,y_i,\underline{\Phi}) \, dy;$$

since for group II, p(m=0) is 0.0 for y scores greater than c,

$$= \sum_{i=1}^{n_2} \ln p(x_i|\underline{\Theta}) + \sum_{i=1}^{n_2} \ln \int_{-\infty}^{c} p(y_i|x_i\,\underline{\Phi}) \, dy.$$

The complete likelihood function is:

$$\ln L = \sum_{i=1}^{n_1+n_2} \ln p(x_i|\underline{\theta}) + \sum_{i=1}^{n_1} \ln p(y_i|x_i, \underline{\theta}) + \sum_{i=1}^{n_2} \ln p(y<c|x_i,\underline{\theta},\underline{\phi}). \quad (26)$$

It should be noted that $\underline{\theta}$ and $\underline{\phi}$ cannot be separated into additive factors and so must be maximized jointly. In this case, where simultaneous maximization must be done, iterative methods are necessary for the calculation of maximum likelihood estimates. Details of the estimation procedure can be found in Maddala (pp. 151-160, 1983).

Another non-ignorable selection process is frequently encountered when there is no observable cut off score on either the x or y variable. In this case, it is useful to think of the selection process as dependent on an unobservable, underlying random variable, u, whose distribution is a function of x and y, and on which there is a threshold value, u_c, marking selection into the sample. The u variable can be viewed as a composite of all

predictor variables, $x_1, x_2, \ldots x_p$ which could include previous achievement, qualifying examinations, one or more admissions interviews, an assessment of motivation, ability to pay, other schools applied to, et.al. Some of these variables may not be observed and/or may not be easily quantified but are nevertheless relevant to the selection process. The observed x variable, whose predictive validity one investigates, typically is but one member of the entire set of selection variables which comprise the u variable. Given x, the probability of selection is the probability that all of the remaining x variables fall in some interval. For example, suppose $u = w_1x_1 + w_2x_2 + w_3x_3$ where x_1 is the observed predictor variable and x_2 and x_3 are unobserved selection variables. Individuals are selected if $u>u_c$. Given x_1, individuals are selected if $(w_2x_2 + w_3x_3) > (u_c - w_1x_1)$. If x_2 and x_3 are related to y, given x_1, it follows that the probability of selection, given x_1, will depend on y and the selection process is non-ignorable. If this is the case, biased parameter estimates of the x y relationship will again result if one simply regresses y on x in the selected group.

In order to model this selection process and use it in a description of the data, i.e., in the likelihood function, we must give mathematical form to our assumptions about the distribution of u given x and y. It is often reasonable to assume that higher scores on the predictor variable, x, are associated with higher scores on y and that, after controlling for x, higher "scores" on u are associated with higher scores on the criterion variable, y. This selection process results in a selected group with individuals who have obtained the higher scores on x and who, given x, are likely to obtain higher scores on y than would have have been obtained by those who were not selected. It should be recalled that since there are no y scores for those excluded, this assumption again cannot be directly confirmed.

A convenient probability model for this type of selection process, assuming that the distribution of u given x and y is normal, is the probit model (Heckman, 1976; Madalla, 1983):

$$p(m=1|x,y) = \int_{-\infty}^{\alpha_0 + \alpha_1 x + \alpha_2 y} \phi = \int_{0}^{\infty} p(u|x,y) \qquad (27)$$

where ϕ is the unit normal density:

$$p(u|x,y) = N(\alpha_0 + \alpha_1 x + \alpha_2 y, 1).$$

It can be seen that the probability of selection is described as a monotonic function of both x and y.

The likelihood function for group 1 with paired x y data and m=1 and for group 2 with x data only and m=0 are as follows:

$$\ln L_1 = \sum_{i=1}^{n_1} \ln p(x_i|\underline{\Theta}) p(y_i|x_i\underline{\Theta}) p(m_i=1|x_i,y_i,\underline{\Phi}) \qquad (28)$$

$$= \sum_{i=1}^{n_1} \ln p(x_i|\Theta) + \sum_{i=1}^{n_1} \ln p(y_i|x_i\underline{\Theta}) p(m_i=1|x_i,y_i,\underline{\Phi})$$

$$\ln L_2 = \sum_{i=1}^{n_2} \ln p(x_i|\underline{\Theta})p(m_i=0|x_i,y_i,\underline{\Phi})$$

$$= \sum_{i=1}^{n_2} \ln p(x_i|\underline{\Theta}) + \sum_{i=1}^{n_2} \ln \int_y p(y_i|x_i,\underline{\Theta})p(m_i=0|x_i,y_i,\underline{\Phi})dy$$

The full log likelihood, $L=L_1+L_2$, is:

$$\ln L = \sum_{i=1}^{n_1+n_2} \ln p(x_i|\underline{\Theta}) + \sum_{i=1}^{n_1} \ln p(y_i|x_i,\underline{\Theta})p(m_i=1|x_i,y_i,\underline{\Phi})$$

$$+ \sum_{i=1}^{n_2} \ln \int_y p(y_i|x_i\underline{\Theta})p(m_i=0|x_i,y_i,\underline{\Phi})dy$$

Again, it should be noted that $\underline{\Theta}$ and $\underline{\Phi}$ cannot be separated into additive factors, and must therefore be maximized jointly. The maximum likelihood estimates can be worked out for all parameters in the model. Details of the procedure can be found in Maddala (1983).

One way of using this model of the selection process was developed by Heckman (1976). Using the probit model, Heckman developed a two stage technique for estimating the

parameters of the full regression model (20). With the two stage technique, it is possible to obtain a set of consistent parameter estimates which closely approximate the maximum likelihood estimates (Heckman, 1976; Müthen and Jöreskog, 1983). Olson and Becker (1983) recognized the importance of the non-ignorable selection process to fields outside econometrics, and recommended in particular that the techniques of the two stage analysis be applied to the field of personnel psychology.

The probit model provides a mathematical description of one selection process, one which often conforms to our assumptions about how the process is working. It should be pointed out, however, that other processes are possible. It might be argued, for example, that the selected group tended to have individuals with middle range x and y scores and that the individuals with either low or high scores were not chosen or were self-selected out. It is clear that such an assumption about the selection process, i.e., about $p(m=1|x, y, \underset{\sim}{\phi})$, would require a different probability model.

It can be said in summary that the work in econometrics has clearly demonstrated the necessity of

modelling the missing data in the case of the non-ignorable selection process. It has also advanced the techniques of probability modelling and maximum likelihood estimation. The cases used to illustrate non-ignorable selection have involved a single predictor, a single criterion, and a single indicator variable. The analysis of a real data set is expected to illuminate the complexities inherent in the approach. Analysis of complex psychometric data using maximum likelihood methods should provide a useful methodological demonstration of the approach in general, as well as an opportunity to explore the application of work done primarily in econometrics to the field of psychometrics.

CHAPTER 3

METHODS

The psychometrician frequently encounters complex data sets with several variables and with data missing on one or more of these variables. The data set which we will use conforms to this description. Using binary variables to indicate selection process, we can represent the data as in Figure 12.

This data set, obtained from a graduate unit of the City University of New York, consists of scores on two predictor variables, undergraduate grade point average (available for all applicants) and Graduate Record Examination (not mandatory and so not available for all applicants), one criterion variable, graduate grade point average (available only for those who enter), and two indicator variables which correspond to whether (m=1) or

not (m=0) the GRE and the GGPA is available (m_y and m_z respectively). Of the 363 who applied, 214 entered for graduate study and obtained a graduate grade point average (groups II and III). Those in groups I and IV did not enter, some because they were not admitted, some because they chose to go elsewhere.

Group	UGGPA(x)	m_y	GRE(y)	m_z	GGPA(z)
I	x	1	y	0	?
	x	1	y	0	?
n=55	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	x	1	y	0	?
II	x	1	y	1	z
	x	1	y	1	z
n=60	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	x	1	y	1	z
III	x	0	?	1	z
	x	0	?	1	z
n=154	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	x	0	?	1	z
IV	x	0	?	0	?
	x	0	?	0	?
	.	.	.	.	.
n=94	.	.	.	.	.
	.	.	.	.	.
	x	0	?	0	?

N=363

Figure 12. Schematic representation of grouped data with number of subjects per group.

Our objective is to estimate the relationships among the variables in order to answer three questions. 1) How well does undergraduate GPA predict graduate school

performance? The parameter of interest is ρ_{xz}. 2) How well does the GRE predict graduate GPA? The parameter of interest is ρ_{yz}. 3) Is prediction improved by using both undergraduate GPA and GRE or could the GRE be ignored? The parameter of interest is equivalent to $\rho^2_{z.xy} - \rho^2_{z.x}$. As we have shown, when there are missing data, answers to these questions are not necessarily straightforward and depend on what selection process is assumed to have led to the data being missing. We can expect different estimates depending on the selection process assumption.

We will answer the three prediction questions concerning ρ_{xz}, ρ_{yz}, and $\rho^2_{z.xy} - \rho^2_{z.x}$ using the maximum likehihood approach and will obtain maximum likelihood estimates of the population parameters under three sets of assumptions concerning the selection processes. In the first analysis, we will assume that both of the selection processes are independent of the x y z data; in the second, that the selection processes are related to x but not related to y or z, given x; and in the third, that only one of the processes is ignorable. The other process we will assume to be non-ignorable, i.e., dependent on a variable with missing data. The differences in the

resulting sets of parameter estimates will illustrate the importance of the selection process assumptions.

The complete likelihood for our data, $L=(x,y,z,m_y,m_z|\ \underline{\Theta},\ \underline{\phi}_y,\ \underline{\phi}_z)$, where $\underline{\Theta}$ contains the parameters underlying the x y z distribution ($\hat{\mu}_x, \hat{\sigma}_x^2, \hat{\beta}_{0y}, \hat{\beta}_{1y}$ $\sigma_{y|x}^2, \hat{\beta}_{0z}, \hat{\beta}_{1z}, \hat{\sigma}_{z|x}^2, \hat{\sigma}_{yz|x}$) and $\underline{\phi}_y$ and $\underline{\phi}_z$ contain the parameters underlying the first and second selection processes, can conveniently be written as the sum of the log likelihoods for each group: $L=L_1+L_2+L_3+L_4$. In the general case, they are as follows:

$$\ln L_1=\sum_{i=1}^{n_1} \ln\ p(x_i,y_i|\ \underline{\Theta})\ +$$

$$\sum_{i=1}^{n_1} \ln\int_z p(m_{yi}=1,m_{zi}=0\,|x_i,y_i,z_i,\underline{\phi}_y\ \underline{\phi}_z)p(z_i\,|\,x_i,y_i,\underline{\Theta})$$

$$\ln L_2=\sum_{i=1}^{n_2} \ln\ p(x_i\,y_i\,z_i|\ \underline{\Theta})+\ \sum_{i=1}^{n_2} \ln\ p(m_{yi}=1,m_{zi}=1|\,x_i,y_i,z_i,\underline{\phi}_y\ \underline{\phi}_z)$$

$$\ln L_3 = \sum_{i=1}^{n_3} \ln p(x_i z_i \mid \underline{\theta}) + \sum_{i=1}^{n_3} \ln \int_y p(m_{yi}=0, m_{zi}=1 \mid x_i, y_i, z_i, \underline{\phi}_y, \underline{\phi}_z)\, p(y_i \mid x_i, z_i, \underline{\theta}) \tag{30}$$

$$\ln L_4 = \sum_{i=1}^{n_4} \ln p(x_i \mid \underline{\theta}) + \sum_{i=1}^{n_4} \ln \iint_{z\,y} p(m_{yi}=0, m_{zi}=0 \mid x_i, y_i, z_i, \underline{\phi}_y, \underline{\phi}_z)\, p(y_i, z_i \mid x_i, \underline{\theta})$$

where n refers to the number of subjects in the group indicated by the n-subscript.

Analysis 1 Ignorable Selection Process

In our first analysis, we assume that both selection processes are ignorable, i.e., independent of y and z, given x:

$$p(m_y m_z \mid x, y, z) = p(m_y, m_z).$$

In this analysis, we assume that the data are missing at

random and not as a function of x, y, or z. In addition we assume that x, y, and z are from a trivariate normal distribution.

Under the assumptions of this analysis, the log liklihood functions can be written as follows:

$$\ln L_1 = \sum_{i=1}^{n_1} \ln p(x_i, y_i | \underline{\Theta}) + \sum_{i=1}^{n_1} \ln p(m_{yi}=1, m_{zi}=0 | \underline{\phi}_y \ \underline{\phi}_z)$$

$$\ln L_2 = \sum_{i=1}^{n_2} \ln p(x_i, y_i, z_i | \underline{\Theta}) + \sum_{i=1}^{n_2} \ln p(m_{yi}=1, m_{zi}=1 | \underline{\phi}_y \ \underline{\phi}_z) \qquad (31)$$

$$\ln L_3 = \sum_{i=1}^{n_3} \ln p(x_i, z_i | \Theta) + \sum_{i=1}^{n_3} \ln p(m_{yi}=0, m_{zi}=1 | \underline{\phi}_y \ \underline{\phi}_z)$$

$$\ln L_4 = \sum_{i=1}^{n_4} \ln p(x_i | \Theta) + \sum_{i=1}^{n_4} \ln p(m_{yi}=0, m_{zi}=0 | \underline{\phi}_y \ \underline{\phi}_z).$$

It can be seen that in this case, $\underline{\Theta}$ and the selection process parameters, $\underline{\phi}_y$ and $\underline{\phi}_z$, are in different additive factors and can therefore be maximized separately. The likelihood, L, can be maximized without regard to the selection processes. Given the added assumption that group

II is a random sample from a trivariate normal x y z distribution, maximization of the likelihood for group II alone will give maximum likelihood estimation of the parameters underlying the entire data set and information from groups I, III, and IV can be discarded. We recognize this to be equivalent to a regression analysis based only on those subjects who have x, y, and z scores. Such a regression analysis is frequently encountered in situations where there are missing data, but the assumptions noted above are rarely made explicit. If we are justified in making these assumptions, the parameters can be estimated as follows:

$\hat{\mu}_x = \bar{x}_2$ =mean of the n_2 x scores

$\hat{\sigma}_x^2 = s_{x2}^2$ =variance of the n_2 x scores

$\hat{\beta}_{oy} = b_{0y}$ =least squares regression constant for predicting y from x, computed from the x y data of the n_2 subjects

$\hat{\beta}_{1y} = b_{1y}$ =least squares regression coefficient for predicting y from x, computed from the x y data of the n_2 subjects

(32)

$\hat{\sigma}_{y|x} = s^2_{y|x}$ =the residual variance computed from the x y data of the n_2 subjects

$\hat{\beta}_{oz} = b_{0z}$ =the least squares regression constant for predicting z from x, computed from the n_2 subjects

$\hat{\beta}_{1z} = b_{1z}$ =the least squares regression coefficient for predicting z from x, computed form the x z data of the n_2 subjects

$\hat{\sigma}^2_{z|x} = s^2_{z|x}$ =the residual variance computed from the x z data of the n_2 subjects

$\hat{\sigma}_{yz|x} = s^2_{yz|x}$ =the residual variance computed from the x y z data of the n_2 subjects.

Again, by replacing the parameters in (13) with these maximum likelihood estimates, we obtain estimates of the population parameters of interest, ρ_{yz}, ρ_{xz}, $\rho^2_{z.xy} - \rho^2_{z.x}$.

Analysis 2 Ignorable Selection Process

Our second analysis is based on the assumption that

both selection processes are ignorable, i.e., independent of y and z, given x, but are both dependent on x:

$$p(m_y, m_z | x, y, z) = p(m_y, m_z | x).$$

In this analysis, we assume that the data are missing as a function of UGPA (x), but independently of GRE (y) score and GGPA (z). This would be the case for m_y, for example, if those with low UGPA tended to feel that they needed a supplementary score to submit with their application and were therefore more likely than those with higher UGPA to take the examination. For m_z, it might be the case if the admissions committee tended to accept those with high undergraduate grades and if, in addition, those who were accepted tended to enter (rather than go elsewhere). In addition, we make the same distribution assumption as in the first analysis, i.e., that x, y, and z are from a trivariate normal distribution.

The likelihood functions in this case can conveniently be arranged as follows:

$$L_1 = \sum_{i=1}^{n_1} \ln p(x_i | \underline{\Theta}) + \sum_{i=1}^{n_1} \ln p(y_i | x_i, \underline{\Theta}) + \sum_{i=1}^{n_1} \ln p(m_{yi}=1, m_{zi}=0 | x_i \underline{\Phi}_y \underline{\Phi}_z)$$

$$L_2 = \sum_{i=1}^{n_2} \ln p(x_i | \underline{\Theta}) + \sum_{i=1}^{n_2} \ln p(y_i z_i | x_i, \underline{\Theta}) + \sum_{i=1}^{n_2} \ln p(m_{yi}=1, m_{zi}=1 | x_i \underline{\Phi}_y \underline{\Phi}_z)$$

$$L_3 = \sum_{i=1}^{n_3} \ln p(x_i | \Theta) + \sum_{i=1}^{n_3} \ln p(z_i | x_i, \underline{\Theta}) + \sum_{i=1}^{n_3} \ln p(m_{yi}=0, m_{zi}=1 | x_i \underline{\Phi}_y \underline{\Phi}_z) \qquad (33)$$

$$L_4 = \sum_{i=1}^{n_4} \ln p(x_i | \Theta) + \sum_{i=1}^{n_4} \ln p(m_{yi}=0, m_{zi}=0 | x_i \underline{\Phi}_y \underline{\Phi}_z)$$

In this case, as in analysis 1, $\underline{\Theta}$ and the selection process parameters are in separate additive functions and

need not be jointly maximized. The likelihood, L, can be maximized without regard to the selection process. Unlike analysis 1, however, the x scores for group 2 cannot be assumed to represent a random sample from the x distribution. As a consequence, in order to estimate the parameters underlying the x y z distribution, one must maximize the entire function ($L = L_1 + L_2 + L_3 + L_4$, excluding the terms with m_y and m_z), necessitating the use of all of the data from each of the groups. In one sense, the analysis could be considered straightforward since the selection processes are assumed to be dependent on a full data variable and ignorable. If the data were in a pattern as shown in Figures 2 - 4, maximum likelihood estimates of the parameters could be obtained by correcting for restriction of range with the appropriate correction formula. However, in this case, the pattern of missing data is more complex and for this reason maximum likelihood estimation is not straightforward.

Analysis 3 Non-ignorable Selection Process

In our third analysis, it is the selection processes (m_y m_z) which must be given mathematical form. We must model, and find ways to mathematically state, our assumptions concerning these processes. In order to do this, we can be guided to some extent by the data, but where there are missing data, we must give form to our subjective beliefs about how the data are missing, i.e., about how the selection processes have worked.

The complete likelihood function for this data set is given by (30). The term in the likelihood function for the selection processes can be conveniently rewritten as follows:

$$p(m_y, m_z \; x,y,z) = p(m_y|x,y,z)p(m_z|m_y,x,y,z).$$

First we consider $p(m_y|x,y,z)$, i.e., the probability of submitting or not submitting GRE scores as it is a function of 1) UGPA (x), 2) a potential or actual score on the GRE (y), and 3) a potential or actual GGPA (z). Since the GRE is optional, only about 30% of all applicants in our sample submitted scores, with total scores ranging from a low of 510 to a high of 1440. The mean was 890.3

with a variance of 378.7. For the 363 applicants, the probability of taking the exam differed at levels of x. Based on this, and with our belief that these applicants are representative of applicants in general, we will assume that the m_y selection process is dependent on x. We hypothesize, however, that m_y is independent of y and z, That is, those who did not take the exam would have had the same distributions of y and z scores as those who did take the exam. It is not evident to us what factors led some to take the exam and others to not take it, but we see no evidence that the factors are dependent on y or z, after controlling for x. We assume, then, that $p(m_y|x,y,z) = p(m_y|x)$.

Next we consider $p(m_z|m_y,x,y,z)$. It should be recalled that $m_z=1$ signifies that an individual entered the graduate program and received a graduate grade point average; where $m_z=0$, it means either that the individual was not accepted or that (s)he was accepted but chose not to enter. First we regard the relationship between m_y and m_z . In our data, there was no relationship found between m_y and m_z after controlling for level of x. This agrees with our belief that the probability of being in the

graduate program had nothing to do with whether or not the GRE was submitted. Neither the admissions committee's decision nor the accepted student's entry decision is likely to have been based on whether or not the GRE was taken. Therefore we will assume that $p(m_z|m_y,x,y,z) = p(m_z|x,y,z)$.

Still considering the conditional probability of m_z, we next look at the relationship between m_z and GRE scores (y). One can ask the question, given knowledge of individuals' x and z scores, would knowledge of their y scores (actual or potential) be likely to improve our prediction of whether or not they are in the graduate program? It is our understanding that at present the GRE score is not considered in the admissions decision. In addition, it seems unlikely that a GRE score alone, i.e., after controlling for undergraduate GPA, would influence an accepted student's decision to enter or not. Therefore we assume that $p(m_z|x,y,z) = p(m_z|x,z)$.

Finally we consider $p(m_z|x,z)$. We believe that a relationship exists between the probability of being in the graduate program and the undergraduate GPA (x) and that this probability is an increasing function of x. Our

data support this conclusion, as does our information that the UGPA is the major variable considered by the admissions committee. We also believe there is likely to be a relationship between the probability of being in the graduate program and graduate GPA (z) after controlling for x. Of those accepted, approximately 30% decide not to enter. We think it likely that this decision is a function of variables which are related to how well the student is likely to perform in graduate school (z). We will assume that the GGPA (z) of those who enter are likely to be higher than the potential GGPA of those who are accepted but do not enter, i.e., that the probability of being in the program is a monotonically increasing function of z, given x.

A summary of the selection model assumptions is as follows:

$$p(m_y, m_z \mid xyz) = p(m_y \mid x)\,p(m_z \mid xz).$$

As in the earlier two analyses, we will assume that x, y, and z are from a trivariate normal distribution.

The likelihood functions for the third analysis are as follows:

$$L_1 = \sum_{i=1}^{n_1} \ln p(x_i|\underline{\Theta}) + \sum_{i=1}^{n_1} \ln p(y_i|x_i\underline{\Theta}) + \sum_{i=1}^{n_1} \ln p(m_{yi}=1|x_i,\underline{\Phi}_y)$$

$$+ \sum_{i=1}^{n_1} \ln \int_z p(z_i|x_i,y_i,\underline{\Theta})p(m_{zi}=0|x_i,z_i,\underline{\Phi}_z)\,dz$$

$$L_2 = \sum_{i=1}^{n_2} \ln p(x_i|\underline{\Theta}) + \sum_{i=1}^{n_2} \ln p(y_i,z_i|x_i,\underline{\Theta}) +$$

$$+ \sum_{i=1}^{n_2} \ln p(m_{yi}=1|x_i\,\underline{\Phi}_y) + \sum_{i=1}^{n_2} \ln p(m_{zi}=1|x_i,z_i,\underline{\Phi}_z) \qquad (35)$$

$$L_3 = \sum_{i=1}^{n_3} \ln p(x_i|\underline{\Theta}) + \sum_{i=1}^{n_3} \ln p(z_i|x_i,\underline{\Theta}) + \sum_{i=1}^{n_3} \ln p(m_{yi}=0|x_i,\underline{\Phi}_y)$$

$$+ \sum_{i=1}^{n_3} \ln p(m_{zi}=1|x_i,z_i,\underline{\Phi}_z)$$

$$L_4 = \sum_{i=1}^{n_4} \ln p(x_i|\underline{\Theta}) + \sum_{i=1}^{n_4} \ln p(m_{yi}=0|x_i,\underline{\Phi}_y)$$

$$+ \sum_{i=1}^{n_4} \ln \int_z p(z_i|x_i,\underline{\Theta})p(m_{zi}=0|x_i,z_i,\underline{\Phi}_z)\,dz$$

In this case, it will be noted that $\underline{\Theta}$ and $\underline{\Phi}_z$ are

under the integral together and not in separate additive factors. As a consequence, the parameters in $\underline{\theta}$ and $\underline{\phi}_2$ must be maximized jointly. In order to do this, we must model the selection process, i.e., provide a mathematical description of the probability of being in the graduate program, given z and x: $p(m_Z|x,z,\underline{\phi}_2)$.

Our assumption concerning the second selection process was that the probability of selection is an increasing monotonic function of x and z. One can view the process which results in subjects being in the institution and being observed on z as being a function of x and additional unmeasured variables. If we assume that these additional variables are monotonically related to z, the probability of selection will be a monotonic function of x and z. We can use the probit model to mathematically describe this selection process for the likelihood function:

$$p(m_Z|x,z,\underline{\phi}_2) = \int_{-\infty}^{\alpha_0 + \alpha_1 x + \alpha_2 z} \phi_2 \qquad (35)$$

From the three analyses, we will have obtained three sets of parameter estimates which we expect to vary depending on which selection process assumption is employed. Using these sets of estimates, we will see how the assumptions lead to different conclusions concerning the relationships among the variables in our data. The model for the third analysis is the most general, and the models of the first and second analyses are nested within it, , i.e., are special cases of the more general model. Because the models are nested in this way, we will be able to use the chi square goodness of fit test to compare the fits of the models. It is expected that the results will show the importance of taking into account the selection process when one validates psychological measures for which complete data are not available.

CHAPTER 4

RESULTS

The means and standard deviations of the variables by group can be found in Table 1. The correlation matrices by group are shown in Tables 2 - 4. The raw data are included in Appendix A.

Table 1
Descriptive Statisics
of data by group

		Statistic					
Group	n	$\bar{x}$	s_x	$\bar{y}$	s_y	$\bar{z}$	s_z
I	55	2.89	.463	895.6	20.22	-	-
II	60	2.87	.483	885.5	18.90	3.33	.425
III	154	2.89	.450	-	-	3.27	.467
IV	94	2.78	.479	-	-	-	-

Table 2

Correlation matrix:
Group I

	x	y
x	1.00	0.48
y		1.00

Table 3

Correlation matrix:
Group II

	x	y	z
x	1.00	0.31	0.42
y		1.00	0.29
z			1.00

Table 4

Correlation matrix:
Group III

	x	z
x	1.00	0.36
z		1.00

Analysis 1

In the first analysis, it was assumed that the selection processes which resulted in missing data were independent of the x, y, and z variables; that is, data were assumed to be missing at random. The probabilistic model which describes the randomly missing data was expressed as follows:

$$p(m_y, m_z \mid x,y,z) = p(m_y, m_z)$$

where m_y is the binary variable standing for whether or not a GRE score (y) was submitted by an applying student and m_z is the binary variable standing for whether or not a graduate grade point average (z) was available for the student. The log likelihood function for all of the data in this case is given in (31).

Under the assumption of randomly missing data, maximum likelihood estimates of the parameters underlying the trivariate normal x y z distribution can be obtained simply by analyzing the data from the subgroup which has complete information on all students, i.e., group II where n=60.

The parameter estimates for the analysis 1 are in the

first column of Table 5. The estimates of the first six parameters ($\sigma_{y|x}$, $\beta_{0y|x}$, $\beta_{1y|x}$, $\sigma_{z|x}$, $\beta_{0z|x}$, $\beta_{1z|x}$) were obtained by regressing y on x, and z on x, in the selected group. Using variances and covariances computed on this group, the conditional covariance ($\sigma_{yz|x}$) was calculated as:

$$\sigma_{yz|x} = s_{yz} - b_{1z}\ b_{1y}\ s_x^2 \qquad (36)$$

Estimates of population correlation coefficients were obtained as follows:

$$\begin{aligned} \hat{\rho}_{xz} &= r_{xz} \\ \hat{\rho}_{yz} &= r_{yz} \end{aligned} \qquad (37)$$

Table 5

Parameter Estimates Obtained as a Function of Selection Process Assumptions

	Analysis		
	1	2	3
Parameter			
$\hat{\sigma}_{y\|x}$	1.81 (0.17)	1.79 (0.12)	1.89
$\hat{\beta}_{0y\|x}$	5.42 (1.43)	4.17 (0.96)	5.30
$\hat{\beta}_{1y\|x}$	1.19 (0.49)	1.63 (0.33)	1.25
$\hat{\sigma}_{z\|x}$	0.39 (0.04)	0.45 (0.02)	0.42 (0.04)
$\hat{\beta}_{0z\|x}$	2.27 (0.31)	2.06 (0.19)	2.15 (0.44)
$\hat{\beta}_{1z\|x}$	0.37 (0.11)	0.42 (0.07)	0.38 (0.09)
$\hat{\sigma}_{z,y\|x}$	0.13	0.18 (0.10)	0.16
$\hat{\rho}_{z.x}$	0.42	0.40	0.39
$\hat{\rho}_{z.y}$	0.29	0.34	0.30
$\hat{\rho}_{z.xy}$	0.45	0.45	0.43

Note: estimated asymptotic standard errors are shown in parentheses.

The estimate for the multiple correlation, $\hat{\rho}_{z.xy}$, is

$$\rho_{z.xy} = \sqrt{\frac{r_{xz}^2 + r_{yz}^2 - 2r_{xy}r_{yz}r_{xz}}{1 - r_{xy}^2}} \qquad (38)$$

The undergraduate grade point average (x), accounting for slightly over .176 ($\hat{\rho}^2_{xz} = .42$) of the variance in the graduate grade point average (z), had better predictive validity than the Graduate Record Examination (y) which accounted for slightly over .084 ($\hat{\rho}^2_{yz} = .29$) of the GGPA variance. Using both predictors resulted in a squared multiple correlation of .203, an increase in predictive power over x alone of 0.03. This increment is not statistically significant at the .05 level.

Analysis 2

In the second analysis, it was assumed that both selection processes were dependent on the undergraduate GPA (x), but independent of the missing data variables, y and z. The probabilistic model describing this selection process is:

$$p(m_y, m_z \mid x, y, z) = p(m_y, m_z \mid x)$$

The log likelihood function for all of the data in this

case is given in (33). In this case, where the data are assumed to be missing in a nonrandom fashion, i.e., as a function of x, it is necessary to maximize the entire likelihood function, excluding the terms with m_y and m_z, in order to obtain maximum likelihood estimates of the parameters underlying the x y z distribution.

In order to obtain these maximum likelihood estimates, a FORTRAN program was written to maximize the log likelihood function as a function of the seven parameters of interest ($\sigma_{y|x}$, $\beta_{0y|x}$, $\beta_{1y|x}$, $\sigma_{z|x}$, $\beta_{0z|x}$, $\beta_{1z|x}$, $\sigma_{yz|x}$), given the entire data set. The subprogram ZXMIN from the IMSL Library (1984) was used to obtain parameter estimates. For the parameter starting values required by the iterative procedure, consistent parameter estimates were obtained from regressions using segments of the data. Data from Groups I and II were combined to produce a segment containing all subjects with both x and y scores. The statistics obtained from regressing y on x in this group, b_{0y} , b_{1y} , $s_{y|x}$ were used as starting values for the parameters, $\beta_{0y|x}$, $\beta_{1y|x}$, $\sigma_{y|x}$. Data from Groups II and III were combined to produce a segment containing all subjects with both x and z scores. The statistics obtained

from regressing z on x in this segment, b_{0z} , b_{1z} , $s_{z|x}$, were used as starting values for the parameters, $\beta_{0z|x}$, $\beta_{1z|x}$, $\sigma_{z|x}$. In order to obtain a starting value for $\sigma_{yz|x}$, group II was used to obtain estimates of s_{xy} , s_{yz} , and s_{xz} ; s_x was obtained using x data from all subjects; the starting value statistic ($s_{yz|x}$) was calculated as

$$s_{yz|x} = s_{yz} - \frac{s_{xy}\, s_{xz}}{s_x^2} \qquad (39)$$

The parameter estimates under the second model can be found in the second column of Table 5. The estimates are substantially the same as those obtained in analysis 1.

Third Analysis

The third analysis was undertaken to study the effect of a selection process which is probabilistically related to the variable on which there are missing data, i.e., the non-ignorable selection process. Our initial examination of the data led us to model the selection process as follows:

$$p(m_y, m_z \mid x,y,z) = p(m_y \mid x)\, p(m_z \mid x,z).$$

Given x, m_y is independent of y and z. In addition, given

x, m_z is independent of m_y and y, but dependent on z. The second selection process was assumed to be a monotonically increasing function of x and z. The probit model was chosen to mathematically describe this selection process:

$$p(m_z|x,z) = \int_{-\infty}^{\alpha_0 + \alpha_1 x + \alpha_2 z} \phi \qquad (40)$$

where ϕ is the unit normal probability density.

We can view the selection process in terms of the following structural model:

$$\begin{aligned} z &= \beta_{0z} + \beta_{1z} x + e \\ z_1 &= v_0 + v_1 x + e_1 \end{aligned} \qquad (41)$$

where, given x, z is normal with mean equal to $(\beta_{0z} + \beta_{1z} x)$ and standard deviation, $\sigma_{z|x}$, and z_1 is normal with mean equal to $v_0 + v_1 x$ and standard deviation, $\sigma_{z_1|x}$, equal to one; z and z_1 are jointly normal with correlation ρ_{zz_1}. The latent variable, z_1, is not directly observable; it is only possible to observe whether or not a score on z_1 is greater than a cutoff score, c, i.e., $z_1 > c$. The event $z_1 > c$ is denoted as $m_z = 1$. The observed variable, z, is

observed if and only if $z_1 > c$.

If we consider the conditional distribution of z_1 given z and x, we can express the probability that $m_Z=1$ as a function of z and x. In this case the model is expressed in terms of a new set of parameters, $\alpha_0, \alpha_1, \alpha_2$.

$$p(m_Z=1 \mid x,z) = p(z_1>c \mid x,z) = \int_{-\infty}^{\alpha_0+\alpha_1 x+\alpha_2 z} \phi \qquad (42)$$

The α's are simply reparameterized versions of the v_i, $\beta_{0z|x}$, $\beta_{1z|x}$, $\sigma_{z|y}$ and $\rho_{z_2 z_1}$ parameters.

In order to obtain maximum likelihood estimates (MLE) of the parameters of the model, $\alpha_0, \alpha_1, \alpha_2, \beta_{0y|x}, \beta_{1y|x}, \sigma_{y|x}$ $\beta_{0z|x}, \beta_{1z|y}, \sigma_{z|x}, \sigma_{y_2|y_1}$, a FORTRAN program was written to maximize the log likelihood function given by (34). Consistent starting values for the parameters were investigated in the following way.

First, given the proposed model, it is straightforward to show

$$p(y|x,z,m_y=m_Z=1) = p(y|x,z)$$

Thus, one can use the xyz data in group II to estimate the

parameters of the conditional distribution of y given x and z, i.e., $\beta_{0y|xz}, \beta_{1y|xz}, \beta_{2y|xz}, \sigma_{y|xz}$. These estimates are obtained simply by regressing y on x and z in group II. It should be noted that although these are the parameters underlying the conditional distribution of y, given x and z, transformation of these parameters into the parameters of interest, i.e., those underlying the distribution of y, given x, can be easily obtained using estimates of the parameters underlying the distribution of z, given x.

Second, one can estimate the selection model parameters, α_0, α_1, α_2, and the parameters for the distribution of z, given x, $\beta_{0z|x}, \beta_{1z|x}, \sigma_{z|x}$. This procedure is based on Heckman's (1976) two stage estimation procedure. The steps can be outlined as follows:

(a) Using the data on all four groups, perform a probit analysis predicting m_Z from x, i.e., estimate

$$v_0 = \frac{\alpha_0 + \alpha_2 \beta_{0z|x}}{(1 + \alpha_1^2 \sigma_{z|x}^2)^{1/2}} \tag{43}$$

$$v_1 = \frac{\alpha_1 + \alpha_2 \beta_{1z|x}}{(1 + \alpha_1^2 \sigma_{z|x}^2)^{1/2}} \tag{44}$$

(b) For each subject in groups II and III, compute the variable

$$x^{*} = \frac{f(v_0 + v_1 x)}{F(v_0 + v_1 x)} \qquad (45)$$

where f= the unit normal probability density,

F= the unit normal distribution function.

(c) Regress z on x_0, x_1, and x^{*} in groups II and III.

(d) Using the results given in steps a-c, one obtains consistent estimates of $\alpha_0, \alpha_1, \alpha_2, \beta_{0z|x}, \beta_{1z|x}, \sigma_{z|x}$.

(e) The estimates obtained in (d) can in turn be used as starting values for calculating MLE of the parameters $\alpha_0, \alpha_1, \alpha_2, \beta_{0z|x}, \beta_{1z|x}, \sigma_{z|x}$. In this analysis, one considers a likelihood function consisting of two types of terms. For subjects in groups II and III, we have $p(z, m_Z=1|x)$. For subjects in groups I and IV, we have $p(m_Z=0|x)$.

Steps a-e were performed using the LIMDEP program (Greene, 1984). Unfortunately, we could not

compute the starting values described in step (e) since the program would not converge. This was in part due to the unreasonable regression parameter estimates obtained in step (d): $\beta_{0z|x} = 13.80$, $\beta_{1z|x} = -.841$, $\beta_{2z|x} = -12.30$. Attempts to use the two stage estimates from (d) and other starting values in the full maximum likelihood procedure also failed because of lack of convergence.

This inability to obtain reasonable starting values using the proposed selection model led us to investigate other models. A model of the following form was considered:

$$p(m_y, m_z \mid x,y,z) = p(m_y \mid x)p(m_z \mid x, m_y, m_y x, z). \qquad (46)$$

This model adds to the original selection process assumption the restriction that m_z is dependent on m_y as well as on x and z. In terms of the structural model representation, the latent selection variable, z_1 , is now defined as follows:

$$z_1 = v_0 + v_1 x + v_2 m_y + v_3 m_y x + e_1 . \qquad (47)$$

This model states that the relationship of the z selection process to x varies as a function of whether or not y is

observed (m_y=1). This notion is supported by the data. Table 6 shows the probability of selection on z as a function of x separately for the m_y=1 and m_y=0 groups. It can be seen that $p(m_z=1 \mid x)$ tends to increase when m_y=0. When m_y=1, no such simple pattern is revealed.

Table 6

Probability of Selection on z
At Levels of m_y and x: $P(m_z=1 \mid x, m_y)$

			m_y= 1		
x intervals	2.00 to 2.39	2.40 to 2.78	2.79 to 3.17	3.18 to 3.56	3.57 to 3.96
N	14	39	32	18	12
P	.50	.59	.42	.54	.50
			m_y= 0		
x intervals	2.00 to 2.38	2.39 to 2.76	2.77 to 3.15	3.16 to 3.53	3.54 to 3.92
N	40	73	69	45	21
P	.46	.63	.68	.46	.79

Using this more complex model (47), steps a-e (described above) were performed. First, the probit model in step (a),

$$p(m_z=1 \mid x, m_y) = p(z_1 > c \mid x, m_y) = \int_{-\infty}^{v_0 + v_1 x + v_2 m_y + v_3 x m_y} \phi \qquad (48)$$

was estimated. Using the Heckman procedure (steps b-d), estimates for $v_i, i=0,1,2,3$, $\beta_{0z|x}$, $\beta_{1z|x}$, $\sigma_{z|x}$, $\rho_{zz_1|x}$ were obtained. These were used as starting values for the maximum likelihood procedure in step (e).

The maximum likelihood estimates obtained for the v_i and $\rho_{zz_1|x}$ parameters of the selection model are given in Table 7. It can be seen that when $m_y=1$, the slope coefficient for x is close to zero (.294 + (-.335) = .041). When $m_y=0$, however, the coefficient (.294) is different from zero at the .09 significance level. It should be noted that the estimated correlation between z and z_1, is negligible and

Table 7

Maximum Likelihood Estimates of Selection Model Parameters in the Non-ignorable Case

Parameter	Estimate	Significance Level
v_0	-.525 (.513)	.30
v_1	.294 (.178)	.09
v_2	.689 (.891)	.44
v_3	-.335 (.306)	.27
ρ_{zz_1}	.160 (.886)	.85

Note: standard errors are shown in parentheses.

insignificant. This result suggests that the missing data process for z is ignorable. The estimates for the parameters underlying the z distribution, given x, $\hat{\beta}_{0z|x}$, $\hat{\beta}_{1z|x}$ $\hat{\sigma}_{z|x}$, and those underlying y, given x, can be found in the third column of Table 5.

An examination of Table 5 shows that the parameter estimates of the third analysis are essentially the same as those of the first and second analyses. This result is consistent with the fact that ρ_{zz_1} is negligible, and the original assumption that the missing data process for y is

ignorable. In general, the undergraduate grade point average (x) was a better predictor of graduate grade point average (z) than the Graduate Record Examination (y), and no significant improvement in prediction is gained by using both x and y.

CHAPTER 5

DISCUSSION

It was our purpose to consider the importance of the selection process in the estimation of the relationships among psychometric variables when there are missing data. In our data set, there were three psychometric variables, i.e., undergraduate grade point average (x), Graduate Record Examination score (y), and graduate grade point average (z), with missing data on y and z, and two selection processes. These processes were represented by two indicator variables m_y and m_z . Using the single data set, we did three analyses, and in each case, modelled a different relationship between the selection process variables and the observed variables, x,y,z, i.e., $p(m_y, m_z | x,y,z)$. Each analysis and model was done under a different set of assumptions concerning the nature of the

selection processes.

Under each set of assumptions, we estimated the parameters underlying the x y z trivariate normal distribution. In this way, we were able to compare the resulting estimates and the conclusions based on the estimates which psychometricians would be led to make. We expected that each analysis would lead to different answers to the questions: how well does undergraduate grade point average predict graduate school performance? how well does the Graduate Record Examination predict graduate school performance? and how much improvement is gained in predictive power when both are used?

In the first analysis, we modelled the selection process under the assumption of randomly missing data, i.e., that the y and z data were missing at random with respect to x, y, and z: $p(m_y,m_z \mid x,y,z)=p(m_y,m_z)$. This assumption is the common though usually tacit assumption of analyses which use only data from subjects with complete data and discard the rest.

In the second analysis, we modelled the selection processes under the assumption that the data were missing on y and z as a function of x: $p(m_y,m_z \mid x,y,z)$ =

$p(m_y, m_z | x)$. This analysis, done using all of the data and an iterative maximum likelihood estimation procedure, is in the spirit of analyses which employ restriction of range correction formulas. This model of the selection processes describes the case where x scores alone are used in the selection decision, or, if there are other variables, they are not related to y and z, given x.

In the third analysis, we considered the case where a selection process may be probabilistically related to the missing data variables, i.e., it may be a non-ignorable case. We began by considering a model which assumed that m was an ignorable selection process dependent only on x and that m was a non-ignorable process dependent on x and z. Because we could not obtain good starting values for this model and therefore could not get maximum likelihood parameter estimates, we turned to a more complex model.

We again modelled the m_y selection process under the assumption that the probability of having a y score is related to x but not related to y and z : $p(m_y | x,y,z,m_z) = p(m_y | x)$. Next, we considered a more complex model for the m_z selection process. It was observed that the probabilistic relationship between x and m_z differed as a

function of m_y . When m_y =0, the probability that m_z =1 was an increasing monotonic function of x. When m_y =1, a different relationship existed. Thus we added to the model the restriction that m_z was probabilistically related to m_y .

In order to explain why the m_y variable should have an effect on the relationship between m_z and x, we consider what distinguishes those who do submit GRE scores from those who do not. It is likely to be the case that those who do submit GRE scores are applying to other, additional schools which require GRE scores of all applicants. For these subjects, even if the institution under study accepts them on the basis of their x scores, there will be some tendency to go elsewhere. In this group, the probability that they are in the graduate school, $p(m_z=1)$, will not be a simple increasing function of x. On the other hand, those who do not submit a GRE score are not as likely to be applying to other schools, and if accepted will tend to enroll. Since the school selects students on the basis of x, the probability that m_z =1 for this group will be a monotonic increasing function of x.

If we take the v_0, v_1, v_2, v_3 estimates from Table 7, the prediction equation for the latent selection variable, z_1 , can be evaluated. For those who submit the GRE scores, it is as follows:

$$z_1 = -.525 + .294x + .689*1 - .335x*1$$
$$= .164 - .041x.$$

The coefficient for x is close to zero. For those who do not submit the GRE scores, it is:

$$z_1 = -.525 + .294x + .689*0 - .335x*0$$
$$= -.525 + .294x.$$

Here it can be seen that x does have predictive power.

The most striking outcome of the three analyses is the similarity of the sets of resulting parameter estimates as shown in Table 5. Based on the estimated population correlations, $\hat{\rho}_{xz}$, $\hat{\rho}_{yz}$, $\hat{\rho}_{z.xy}$, the psychometricians' conclusions concerning predictive validity would be the same, regardless of the different assumptions of each analysis.

It will be recalled that the methods of the first analysis lead to consistent estimates of the correlation coefficients when 1) the distribution assumption is

satisfied, 2) the selection assumption of ignorablility is satisfied, and 3) the range restrictions do not result in significantly different estimates of the variance of the full data predictor variable. The second analysis was less restrictive than the first in that unbiased estimates of the correlation coefficients would be obtained even if violations of the restriction of range assumption were present. The third analysis was the least restrictive, allowing violation of the restriction of range assumption and allowing for what we considered to be a more realistic model of one of the selection processes. It can be seen (Table 5) that the most restrictive analysis results in essentially the same parameter estimates as do the more complex analyses, and on the grounds of parsimony, this model must be considered the best of the three. As a consequence, we are led to conclude that the selection assumption and the restriction of range assumption are not violated in this data set. (Since all three analyses make the same distribution assumption, the merit of this assumption cannot be weighed here.)

With regard to restriction of range assumption, we see that the variance of x in the restricted group, s_r^2 ,

and in the total group, s_t^2 , were essentially the same: s_r^2 = .23, s_t^2 = .22. Even though x and m_z are related when m_y=1, no such relationship exists for the total group. Therefore, the selected group where m_y=1, m_z=1 can be viewed as a random sample from the population with regard to x. Using all of the x scores and applying the restriction of range correction formulas (1) and (2) would result in the same parameter estimates as those obtained using only the selected group.

With regard to the selection process assumption, it would appear that the missing data are the result of some selection process or processes independent of the x, y, or z variables. There are many factors related to whether or not an individual is in the graduate school. These include motivation, past performance, occupation, ability, and, inevitably, other unmeasureable factors of varying importance. Contrary to our expectations, in the case of this data set, they apparently have together resulted in an ignorable, random selection process. The data behave as if they are missing at random with regard to y and z, as well as x.

This conclusion in regard to the m_z selection

process is further supported by considering the estimated correlation between the latent selection variable, z_1 , and z, as shown in Table 7; $\hat{\rho}_{zz_1|x}$ is not significantly different from zero. The m_z selection process is ignorable, i.e., not related to z.

In short, where we expected three different sets of parameter estimates from the three analyses, we obtained three very similar sets. Since selection was not significantly related to the undergraduate grade point average, there were no significant differences between Analyses 1 and 2. Since selection was not significantly related to the missing data variables, there were no significant differences between Analysis 3 and the other analyses. Contrary to our expectations, the sample with data on all three variables behaves as a random sample from the xyz distribution.

We have seen that disregard of the operation of the selection processes in this data set would not have led to biased parameter estimates or misguided conclusions about predictive validity. Nevertheless, it would rarely be safe to assume that the selection processes are irrelevant in a missing data situation, and ways must be found to take

them into account. Selection process modelling combined with maximum likelihood estimation procedures still seems to be a potentially useful tool, but difficulties with the approach must be pointed out.

Most data analysis procedures may be relied on to produce results which can then be judged in terms of the assumptions of the analysis. The selection process modelling approach, however, will in some cases simply not function; the iterative estimation procedure will not reach convergence. This may be the result of one or more of the following general problems: poor starting values, incorrect modelling, a too small sample size, too many parameters to be estimated, and matters of a variable's limited range and colinearity. When the procedure does not work, it is difficult to determine the cause of the failure and to proceed with correction. The availability of the LIMDEP program enabled us to conclude our work with a modified approach. Had it not been available, reaching a conclusion (if possible) would have been time consuming and costly in terms of computer expense.

While the problems of this maximum likelihood estimation procedure and of selection process modelling

have proved complex, the approach can nevertheless be useful in the analysis of real data sets with missing data. It would be useful to conduct further studies concerning characteristics of data sets which produce difficulties in the maximum likelihood estimation procedure.

Appendix A

Data Set

x	y	z	m_z	m_y
2.72	980		0	1
3.66	850		0	1
2.31	900		0	1
3.23	1100		0	1
3.62	1300		0	1
2.89	710		0	1
2.31	940		0	1
2.25	950		0	1
3.72	780		0	1
2.85	1160		0	1
3.75	1070		0	1
2.33	650		0	1
2.22	510		0	1
2.84	900		0	1
3.08	770		0	1
3.05	850		0	1
2.84	790		0	1
2.87	1040		0	1
2.00	550		0	1
2.74	1080		0	1
3.53	1110		0	1
2.84	600		0	1
3.48	1000		0	1
3.01	1020		0	1
2.47	990		0	1
2.42	940		0	1
3.75	1270		0	1
2.73	970		0	1
2.00	540		0	1
3.18	830		0	1
2.59	920		0	1
2.51	540		0	1
3.17	800		0	1
2.68	690		0	1
2.83	870		0	1
3.00	1360		0	1
3.01	1050		0	1
3.17	620		0	1
2.82	850		0	1
3.26	1080		0	1
2.50	1110		0	1
2.47	570		0	1

2.60	740		0	1
3.04	870		0	1
2.95	950		0	1
3.34	760		0	1
2.40	850		0	1
2.43	610		0	1
2.82	1140		0	1
3.41	1130		0	1
3.34	1020		0	1
3.80	960		0	1
2.77	890		0	1
3.15	990		0	1
2.41	740		0	1

3.27	2.94	1	0
2.20	3.11	1	0
2.62	3.40	1	0
3.71	3.71	1	0
3.62	3.60	1	0
2.44	3.60	1	0
2.83	3.21	1	0
2.47	2.70	1	0
2.14	1.80	1	0
2.70	3.08	1	0
2.98	2.78	1	0
2.99	3.74	1	0
2.33	3.15	1	0
2.84	3.30	1	0
2.60	3.20	1	0
2.68	3.32	1	0
3.0	3.00	1	0
3.06	3.15	1	0
2.44	2.80	1	0
3.24	3.37	1	0
2.95	4.00	1	0
3.75	3.46	1	0
3.15	2.80	1	0
3.00	3.66	1	0
3.00	3.70	1	0
3.89	4.00	1	0
2.50	3.35	1	0
2.96	3.82	1	0
3.45	3.60	1	0
2.70	1.42	1	0

2.12	3.46	1	0	
3.18	2.73	1	0	
2.63	2.56	1	0	
2.07	2.92	1	0	
2.10	2.90	1	0	
2.05	2.75	1	0	
2.70	2.50	1	0	
3.07	3.33	1	0	
2.30	3.30	1	0	
2.95	3.50	1	0	
2.14	2.56	1	0	
2.44	2.91	1	0	
2.21	3.00	1	0	
2.64	3.01	1	0	
2.61	2.46	1	0	
2.88	2.35	1	0	
2.92	3.46	1	0	
2.75	2.82	1	0	
2.71	3.52	1	0	
2.72	3.12	1	0	
2.75	3.10	1	0	
3.49	3.70	1	0	
3.90	3.87	1	0	
3.23	3.61	1	0	
2.36	2.82	1	0	
2.75	2.98	1	0	
3.75	2.65	1	0	
2.87	3.82	1	0	
3.87	3.78	1	0	
3.55	3.76	1	0	
2.33	3.21	1	0	
2.50	3.74	1	0	
2.95	3.34	1	0	
2.76	3.73	1	0	
2.29	3.37	1	0	
2.48	2.35	1	0	
2.41	3.20	1	0	
3.21	3.15	1	0	
2.76	2.95	1	0	
3.06	3.51	1	0	
2.58	3.97	1	0	
3.52	4.00	1	0	
3.65	3.50	1	0	
2.66	3.00	1	0	

2.45	2.96	1	0
2.61	3.63	1	0
2.23	2.90	1	0
2.95	3.07	1	0
3.77	3.40	1	0
2.09	3.03	1	0
2.68	3.90	1	0
3.54	3.45	1	0
3.10	3.88	1	0
3.63	3.53	1	0
2.98	2.97	1	0
2.36	3.26	1	0
2.46	2.94	1	0
2.60	3.18	1	0
2.77	3.18	1	0
2.85	3.78	1	0
2.97	3.00	1	0
2.42	3.00	1	0
2.95	3.00	1	0
3.40	3.44	1	0
3.00	3.25	1	0
2.80	3.41	1	0
3.15	3.78	1	0
2.95	3.31	1	0
3.32	3.28	1	0
2.90	3.92	1	0
2.94	2.30	1	0
3.63	3.24	1	0
3.92	3.92	1	0
3.07	3.52	1	0
3.58	3.56	1	0
2.57	2.99	1	0
2.54	3.34	1	0
2.64	3.66	1	0
3.72	3.86	1	0
2.63	4.00	1	0
2.80	3.65	1	0
3.13	3.34	1	0
2.89	3.15	1	0
2.49	3.35	1	0
2.80	2.98	1	0
3.63	3.59	1	0
2.99	2.20	1	0
3.39	3.32	1	0

2.61		3.48	1	0
2.89		3.31	1	0
2.40		3.02	1	0
2.67		3.50	1	0
3.43		1.50	1	0
3.38		3.30	1	0
3.36		3.34	1	0
2.90		3.82	1	0
2.16		3.11	1	0
2.96		3.84	1	0
2.60		3.13	1	0
2.90		3.68	1	0
2.40		3.58	1	0
2.83		3.67	1	0
3.02		3.38	1	0
2.26		2.74	1	0
3.00		3.82	1	0
3.16		3.72	1	0
2.42		3.71	1	0
2.47		2.98	1	0
2.58		3.61	1	0
3.23		3.34	1	0
2.86		2.75	1	0
3.01		3.27	1	0
2.47		2.66	1	0
2.98		3.07	1	0
3.29		3.58	1	0
2.43		2.80	1	0
3.37		3.95	1	0
3.23		3.5	1	0
2.62		3.28	1	0
3.04		3.55	1	0
2.99		3.55	1	0
3.86		3.70	1	0
2.80		3.47	1	0
3.21		3.72	1	0
2.66	840	3.30	1	1
2.75	630	2.80	1	1
3.96	960	3.96	1	1
3.63	850	3.74	1	1
2.97	740	2.24	1	1
2.50	710	3.30	1	1
3.26	630	3.05	1	1

3.19	820	3.74	1	1
2.25	860	3.37	1	1
2.76	1080	3.67	1	1
3.00	980	3.85	1	1
3.12	810	3.07	1	1
3.44	1140	3.00	1	1
3.61	1290	3.84	1	1
2.50	700	2.65	1	1
3.52	940	3.70	1	1
2.50	850	3.67	1	1
3.05	670	2.64	1	1
3.43	1060	3.77	1	1
2.72	730	3.53	1	1
2.45	740	3.20	1	1
2.52	910	3.78	1	1
2.80	1060	3.61	1	1
2.46	730	3.56	1	1
2.97	780	3.30	1	1
3.38	890	3.37	1	1
2.56	1140	2.76	1	1
2.50	930	3.54	1	1
2.67	990	3.20	1	1
2.40	1000	3.15	1	1
3.31	990	2.66	1	1
2.49	1000	2.76	1	1
3.00	730	3.64	1	1
3.84	930	3.82	1	1
2.04	1000	2.00	1	1
2.59	1080	3.96	1	1
2.42	980	3.26	1	1
3.50	1440	4.00	1	1
2.90	820	3.11	1	1
2.50	990	3.52	1	1
2.52	860	3.23	1	1
2.33	580	3.33	1	1
2.36	720	2.8	1	1
3.85	930	3.45	1	1
2.65	700	3.32	1	1
3.55	950	3.76	1	1
2.52	700	3.06	1	1
2.55	650	3.0	1	1
2.39	740	3.22	1	1
2.23	800	2.95	1	1
3.12	720	3.50	1	1

2.00	890	3.20	1	1
2.54	730	3.19	1	1
3.42	900	3.82	1	1
3.05	780	3.80	1	1
2.98	1440	3.62	1	1
2.60	910	3.37	1	1
3.17	860	3.08	1	1
3.66	1260	3.27	1	1
3.09	590	3.50	1	1
2.38			0	0
2.86			0	0
2.41			0	0
2.82			0	0
2.62			0	0
2.84			0	0
2.28			0	0
2.17			0	0
2.50			0	0
3.41			0	0
3.31			0	0
2.18			0	0
2.42			0	0
3.04			0	0
2.66			0	0
3.19			0	0
2.42			0	0
2.04			0	0
2.00			0	0
3.32			0	0
3.11			0	0
3.31			0	0
3.23			0	0
2.98			0	0
2.17			0	0
3.48			0	0
2.66			0	0
3.49			0	0
2.70			0	0
2.77			0	0
3.01			0	0
2.10			0	0
2.23			0	0
2.13			0	0

2.54	0 0
2.93	0 0
3.49	0 0
3.65	0 0
2.96	0 0
2.75	0 0
3.34	0 0
2.65	0 0
3.22	0 0
3.21	0 0
3.37	0 0
3.24	0 0
2.03	0 0
3.11	0 0
2.21	0 0
3.30	0 0
2.58	0 0
3.31	0 0
3.50	0 0
2.72	0 0
2.86	0 0
3.23	0 0
3.00	0 0
3.20	0 0
3.35	0 0
3.32	0 0
2.67	0 0
3.62	0 0
2.55	0 0
2.12	0 0
2.08	0 0
2.09	0 0
2.70	0 0
2.57	0 0
2.42	0 0
2.65	0 0
2.64	0 0
2.25	0 0
3.41	0 0
2.68	0 0
2.15	0 0
3.01	0 0
2.88	0 0
2.60	0 0

2.94	0 0
2.19	0 0
2.78	0 0
3.20	0 0
3.80	0 0
2.50	0 0
2.92	0 0
2.00	0 0
2.00	0 0
2.77	0 0
3.35	0 0
3.40	0 0
2.65	0 0
2.11	0 0
2.75	0 0
2.00	0 0

BIBLIOGRAPHY

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

Forsyth, R. A. An empirical note on correlation coefficients corrected for restriction of range. Educational and Psychological Measurement, 1971, 31, 115-123.

Greene, W. H. LIMDEP Manual. New York: New York University, 1984.

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

Greener, J. M. & Osburn, H. G. Accuracy of corrections for restriction of range due to explicit selection in heteroscedastic and non-linear distributions. Educational and Psychological Measurement, 1980, 40, 337-345.

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

Heckman, J. J. The common structure of statistical models of truncation, sample selection and limited dependent variables and simple estimation for such models. Annals of Economic and Social Measurement, 1976, 5, 475-490.

Hogg, R. V. & Craig, A. T. Introduction to Mathematical Statistics. New York: McMillan, 1970.

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

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

Linn, R. L., Harnisch, D. L., & Dunbar, S. B. Correction for range restriction; an empirical investigation of conditions resulting in conservative correction. Journal of Applied Psychology, 1981, 66, 655-663.

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

Maddala, G. S. Limited-Dependent and Qualitative Variables in Econometrics. New York: Cambridge University Press, 1983.

Müthen, B. & Jöreskog, K. Selectivity problems in quasi-experimental studies. Paper presented at a conference on experimental research in social sciences. University of Florida, 1981.

Novick, M. R. & Thayer, D. T. An Investigation of the Accuracy of the Pearson Selection Formulas.[Research Memorandum RM69-22]. Princeton, N.J.: Educational Testing Service, 1969.

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

Pearson, K. Mathematical contributions to the theory of evolution XI. On the influence of natural selection on the variability and correction of organs. Philosophical Transactions of the Royal Society of London, Series A, 1903, 200, 1-66.

Rubin, D. B. Inference and missing data. Biometrika, 1976, 63, 581-593.

Tobin, J. Estimation of relationships for limited dependent variables. Econometrica, 1958, 26, 24-36.