DOCUMENT RESUME

ED 359 193 TM 019 813

AUTHOR	Gross, Alan L.; And Others
TITLE	A Model for Investigating Predictive Validity at Highly Selective Institutions.
PUB DATE	Apr 93
NOTE	33p.; Paper presented at the Annual Meeting of the American Educational Re earch Association (Atlanta, GA, April 11-16, 1993).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Admission (School); Algorithms; *Competitive Selection; Higher Education; *Mathematical Models; *Predictive Validity; Secondary Education; Selective Admission
IDENTIFIERS	*Missing Data

ABSTRACT

A statistical model for investigating predictive validity at highly selective institutions is described. When the selection ratio is small, one must typically deal with a data set containing relatively large amounts of missing data on both criterion and predictor variables. Standard statistical approaches are based on the strong assumption that the missing data are missing at random (MAR) (i.e., the missing data can be accounted for in terms of the observed measures), and there are no unmeasured variables that underlie the missing data process. The proposed model represents an attempt to account for any unmeasured selection variables by assuming that applicants are first placed into admission categories by the institution and then selected within each category in terms of the observed predictor variables. Thus, although the MAR assumption may not hold for the set of all applicants, it may very well hold within each admission category. The model uses the EM algorithm to obtain estimates of validity separately within each category. The model is quite general and can be used when there are missing data on the predictor and criterion variables, and even if the admission category is not known for each applicant. The proposed model is illustrated in terms of a real life data set for a selective secondary school with over 2,000 applicants. Four tables present analysis data. (Author/SLD)

A Model for Investigating Predictive Validity at Highly Selective Institutions

Alan L. Gross Department of Educational Psychology, Graduate Center, City University of New York

Rena Subotnik Department of Educational Foundations, Hunter College, City University of New York

Perry Halkitis Department of Educational Psychology, Graduate Center, City University of New York

Richard Klutch Hunter College, City University of New York

Paper Presented at the American Educational Research Association, April, 1993.

Abstract

A statistical model for investigating predictive validity at highly selective institutions is described. When the selection ratio is small, one must typically deal with a data set containing relatively large amounts of missing data on both the criterion and predictor variables. Standard statistical approaches are based on the strong assumption that the missing data are missing at random (MAR), i.e., the missing data can be accounted for in terms of the observed measures, and there are no unmeasured variables which underlie the missing data process. It is well known that violations in this assumption can yield biased estimates especially when there is a high proportion of missing data. The proposed model represents an attempt to account for any unmeasured selection variables by assuming that applicants are first placed into admission categories by the institution and then selected within each category in terms of the observed predictor variables. Thus, although the MAR assumption may not hold for the set of all applicants, it may very well hold within each admission category. The model uses the EM algorithm to obtain estimates of validity separately within each category. The model is quite general and can be used when there are missing data on the predictor and criterion variables, and even if the admission category is not known for each applicant. The proposed model is illustrated in terms of a real life data set.

INTRODUCTION

The problem of investigating the validity of a selection procedure is particularly difficult for highly selective institutions due to the relatively large amount of missing data that typically occurs in this setting. Our original interest in this problem arose because of the need to investigate the validity of a real life admissions program at an educational institution which admits less than 10 percent of their applicants. The basic problem is one of investigating the statistical relationship of a set of predictor variables (x_1, x_2, ..., x_p) to some criterion variable y when complete data cannot be obtained. It should be noted that in the most general case, data will be missing for both the y and x variables. The parameters to be estimated may include the squared multiple correlation in predicting y from the x's, the regression weights, various squared semi-partial correlations showing the "importance" of various x variables in predicting y, and the difference in the expected y score of individuals selected using the x's and those who would have been selected by a lottery. This last parameter is especially relevant for highly selective institutions where it is often argued that the applicants are a self-selected group; thus a randomly chosen group of these applicants will perform at a comparable level to a group selected on the basis of test scores.

How can the observed data be used to estimate the parameters of interest? The answer to this question depends upon what assumptions can be made concerning the selection process which produced the missing data. Whereas these assumptions may have a minor effect on the analysis when the proportion of missing data is small, these assumptions must be carefully considered when there are relatively large amounts of missing data. The simplest (although often not realistic) assumption is that the missing data can be explained in terms of the observed predictor variables. For example, suppose only subjects who score highest on x_1 (an aptitude test) are observed on x_2 (a -tructured interview), and only those subjects scoring highest on x_2 are observed on y. In this example, x_2 is missing as a function of x_1, and y is missing as a function of x_2. When the missing data can be explained simply in terms of the observed predictor variables, the missing data can be described as missing at random (MAR), Little & Rubin (1987). Standard methods of estimation can be applied given the MAR assumption. For example, Little & Rubin (1987), Rubin (1976) have shown that maximum likelihood estimates can be obtained by simply considering the likelihood for the observed data. It should be noted, that the standard correlation corrected for "restriction in range," is the maximum likelihood estimate of the population correlation when there is a single x variable and y is MAR (Cohen, 1955) Unfortunately, in many cases the observed predictor variables alone will not explain the missing data, implying that additional unmeasured variables are operating. If these unaccounted for variables are statistically related to the missing data, given the observed data, the MAR assumption will not hold, and the estimation procedure can become quite complex since one must introduce a statistical model for the unmeasured variables. Failure to account for the unmeasured variables will in general lead to biased results (Heckman, 1976, 1979; Linn, 1968; Olson & Becker, 1983). In our study of the real life data set previously described, we were confronted with just such a problem; a simple inspection of the applicant x scores suggested that the missing data could not be accounted for simply in terms of the observed predictor variables. Many instances were found where non-accepted applicants had higher scores on the observed predictor variables than accepted applicants. Thus, the missing data could not be assumed to be MAR. However, a careful analysis of the problem showed that the unmeasured selection variables could be accounted for by considering a set of admission categories used by the institution. Based on a set of measured "background variables," applicants were first classified by the institution into one of three admission categories. The use of these categories basically represented a policy decision by the institution. For each category, selection was then based on the x variables. Thus, although the MAR assumption does not hold for the data set as a whole, the assumption is tenable within each category. Given this structure where applicants are first grouped into categories and then selected in terms of the x variables, one can use the observed data to estimate the validity of the selection process separately within each admission category. This strategy was adopted as a model for investigating the validity of the admissions procedure. The proposed model is quite general and can be employed when there are missing values on both the x and y variables, and in addition, even if the admission category membership is unknown for some applicants.

In section 2, the general statistical model is described together with the methods for estimating the validity of the predictor variables within each admission category. In section 3 the application of the model to the previously mentioned data set is described. Although we illustrate the usefulness of the proposed model in terms of a single real life data set, we believe that the model can be applied in validation studies at other highly selective institutions. The model is also not limited to educational settings, but could be used in industrial selection programs where different admission standards are used for different well defined applicant groups. In this third section we also deal with the problem of accounting for the missing y scores of admitted applicants who decline the admissions offer or who drop out prior to the measurement of y. For these cases, the missing data may be a function of unmeasured variables which are statistically related to y.

Thus, these missing y scores cannot always be assumed to be MAR. We deal with this issue by performing a sensitivity analysis where different possible values are imputed for the missing scores, and the changes in the final analysis are noted. The last section contains our conclusions and a discussion of future research needs.

2. Statistical Model

We assume that the applicants are grouped into a set of $g > 1$ admission categories. Within each category, the p x variables and the y variable are assumed to be multivariate normal with mean vector μ_j and variance covariance matrix Σ_j, $j = 1,2,...g$. In the ideal case where there are no missing data, one can choose between two different methods for estimating the unknown parameters. First, if one assumes that the Σ_j matrices vary over the categories, the data from each admission category would be separately analyzed. For example, g separate multiple correlations would be computed. Second, if one assumes homogeneity for the Σ_j matrices, a single pooled sample variance covariance matrix is computed. In this case a single multiple correlation would be obtained. Assuming reasonably large sample sizes within each category, the choice between these two options would be based on some common homogeneity of variance-covariance test.

Consider the case of missing data. If category membership is known for all applicants, but there are missing data on the x and y variables (assumed to be MAR within a category), the two options described above will be available as long as there are enough complete cases within each category to assure that the individual Σ_j matrices can be estimated. The choice between the two models (common Σ versus separate Σ_j) can be made using an appropriate statistical test. Due to the missing data, the commonly used homogeneity tests cannot be applied. However, one could test the null hypothesis of equal Σ_j using a log-likelihood ratio test. (Mood, Graybill, & Boes, 1974).

Given that missing data are present, the general method used to obtain estimates in both the homogeneous and heterogeneous Σ_j models is the so called EM algorithm (Little & Rubin, 1987). The procedure can be described in general using a type of "what if" reasoning. If there were complete data, one would compute the usual statistics for estimating the parameters of interest. For example, one would compute the mean of the x and y variables within each category as well as the category variance covariance matrix or the pooled variance covariance matrix. These statistics would represent the maximum likelihood estimates. In the presence of missing data, one estimates these statistics by computing their expected values, given the observed data, and some initial set of parameter estimates. These expected values are then used as a new set of parameter estimates, and the expectation step is again applied. The procedure is continued until convergence is reached. The final expected values are the maximum likelihood estimates. We know of two computer programs for performing this analysis; BMDP-AM (1990) for the unequal Σ_j case, and a FORTRAN program obtained from M.D. Schlucter (personal communication, 1990) for the equal Σ_j case. It should be noted that the latter program can also be used in the most complex case where data are missing for category membership, as well as for the x and y variables. In other words, not only are some subjects missing scores on x and y, but the category membership may not be known for all applicants. In this case, parameter estimation is possible only if the common Σ model is assumed. The EM estimation theory for this case is more complex and is described in detail in Little & Rubin (1987).

To further illustrate the estimation procedure, consider the following simple data set consisting of g=2 admission categories, two x variables (x_1, x_2) and y. Suppose there are $n_1 = 6$ applicants in category 1 and $n_2 = 8$ applicants in category 2. The observed and missing data (denoted as "?") are given as follows:

Category 1

x1	x2	y
5	?	9
3	4	7
2	3	5
4	5	?
?	4	?
?	3	7

Category 2

x1	x2	y
8	7	9
8	9	6
7	6	8
5	5	?
4	6	?
3	2	?
3	?	?
2	?	?

The missing data pattern in category 2 is said to be monotonic or nested since there is an ordering of the variables in terms of "observability." The x_1 variable is most observed, followed by x_2, and then y which is the least observed. The data in category 1 does not exhibit a monotonic pattern.

Let us assume that the 3 by 3 variance covariance matrix of the x's and y's is the same for both categories. (The algorithm follows the same logic for the unequal Σ_j case). The problem is to estimate the means for the variables for each category (μ_{1j}, μ_{2j}, μ_{yj}, j = 1,2) and the common 3 by 3 variance covariance matrix, Σ. Given estimates of these basic parameters, one can readily estimate additional parameters. For example, since the multiple correlation can be computed from the variance covariance matrix, given an estimate of the latter, we can estimate the former. In addition, given an estimate for Σ, one can readily obtain the maximum likelihood estimates for the regression weights in predicting y from x_1 and x_2.

If the data were complete, the maximum likelihood estimates would be obtained by first computing the following sums and sums of squares and cross products for the data in category j, j=1,2:

$$\Sigma x_{1j},\ \Sigma x_{2j},\ \Sigma y_j,$$

$$\Sigma x^2{}_{1j},\ \Sigma x^2{}_{2j},\ \Sigma y^2{}_j,$$

$$\Sigma(x_{1j})(y_j),\ \Sigma(x_{2j})(y_j),\ \Sigma(x_{1j})(x_{2j})$$

The traditional maximum likelihood estimates are then obtained by computing the means from the sums, and transforming the sums of squares and cross products into variances and covariances.

When missing data are present, these statistics are iteratively estimated by the EM algorithm. Consider as an example the estimation of the sum of the y's in category 1. Suppose initial estimates of the unknown parameters (μ_1, μ_2, and Σ) are obtained from the complete data. For example, the mean of y in category 1 is initially estimated to be 28/4 = 7. To estimate the sum of y in category 1, the missing y scores of cases 4 and 5 are estimated. These estimates are the expected values given the observed data and the initial estimates. For subject 4, y is estimated from the regression equation predicting y from x_1 and x_2. These regression weights are in turn obtained from the initial estimate for Σ. Similarly, for subject 5, y is estimated from the regression equation predicting y from x_2. Similar computations are followed to estimate all of the sums and sums of squares of cross products. For example to estimate the sum of squares of y in category 1 ($\Sigma y^2{}_1$) the missing values for $y_4{}^2$ and $y_5{}^2$ are replaced with their expected values. In this case the estimates are given by the sum of ε squared predicted value and a residual variance. Once all of the expected values are computed, the usual maximum likelihood estimates are computed and used as new parameter estimates. The expectation step is then performed again. The process continues until the estimates converge. As previously noted, computer programs are available for performing this analysis.

It should be noted that in category 2 where the missing data pattern is monotonic, the maximum likelihood estimates can be obtained in a simple non-iterative manner without using the EM algorithm. The estimates of the mean and variance of x_1 can be directly computed since there are no missing values for this variable. Secondly, two regression equations and associated residual variances can be computed predicting x_2 from x_1, and y from x_1 and x_2. The latter two analyses are performed using the complete x_1,x_2 and the complete x_1,x_2,y data sets respectively. The computed statistics all provide maximum likelihood estimates under the MAR assumption. Using these estimates, one can readily compute the maximum likelihood estimates of any other parameters which are expressible in terms of the originally estimated parameters. For example, the estimate of the mean of x_2 can be obtained by evaluating the regression equation predicting x_2 from x_1 at the estimated mean for x_1. Similarly, the formula for the population multiple correlation of y with x_1 and x_2 can be expressed as a function of the variance of x_1 and the regression weights and residual variances for predicting x_2 from x_1 and y from x_1 and x_2. Replacing these regression weights and variances by their estimates yields the maximum likelihood estimate of the multiple correlation. The basis for these simplified non-iterative analyses when the missing data pattern is monotonic is explained in Little & Rubin (1987). It is interesting to note that if this logic is applied in estimating the xy correlation for the very special monotonic pattern which arises when there is only a single x variable measured on all applicants but there are missing data on y, the resulting estimate is the common restriction in range correction formula. In general, one can estimate the xy correlation when there are missing data on both x and y (a non-monotonic pattern). Further, one can estimate the multiple correlation in the case where there are several x variables, and the missing data for the x's and y are not monotonic. In these last two examples, the iterative EM algorithm provides the general method for obtaining the estimates. Simple formulas such as the correction formula do not exist in these cases.

3. Application to Real Data

The model described above was applied in an investigation of the predictive validity of the selection process at a prestigious Northeastern secondary school where admission is based on the performance of applicants on a three part examination consisting of Mathematics, English, and Essay sections. Each of the applicants to the institution is first categorized into one of three admission categories based on background variables which include SES factors and educational history. Within each of these categories, the selection procedure involves two steps. First, applicant examinations are scored on both the Mathematics and English sections and those students whose scores fall below a determined cutoff are eliminated. The essay of the remaining applicants are then read and scored and those applicants receiving a passing score on the Essay are invited to attend the school. It should be noted that different cutoff scores are used within each admission category. The criterion or y variable (second year grade point average, GPA) is measured for nearly all invited applicants. Less than ten percent of the invited applicants decline the offer or leave the school prior to the measurement of GPA. In summary, within each admission category there are no missing data values for the Mathematics and English examinations, and the missing data pattern for the Essay and GPA variables is monotonic.

As noted above, the GPA variable is measured on over 90 percent of the cases who survived the second stage of selection. More specifically, there were missing GPA scores for a total of 18 individuals who were offered admission but declined or dropped out prior to the time that GPA was recorded. If it were not for these 18 cases, it is clear that the missing data within each of the three admission categories could be assumed to be MAR (Essay scores are missing as a function of Mathematics and English scores; GPA is missing as a function of Mathematics, English, and Essay scores). In the analyses to be described below, we will in fact treat these 18 cases as if they were MAR, i.e, as if they were simply not admitted on the basis of the three predictor variables. The possible bias introduced by this assumption will be later investigated by performing a sensitivity analysis where a range of values for the 18 Y scores are imputed and the parameter estimates recomputed.

The observed characteristics of the applicants (means, standard deviations, group sizes) for the 1988-1989 academic year are shown in Table 1. The descriptive statistics are reported separately for each admission category.

INSERT TABLE 1 ABOUT HERE

It is seen in Table 1 that every one of the 44 applicants in category 1 is accepted. The only missing data occur on the GPA variable for a single student who declined admission. It should be noted that even though all category 1 applicants were accepted for the 1988-1989 school year, it was still of interest to consider the predictive validity of the admissions tests since in subsequent years it is quite possible that the selection ratio for category 1 may be less than 1.00 In category 2, there were 1845 applicants of whom only 365 survived the first selection stage and were measured on the Essay, and from this latter group, only 151 were offered admission. Of these 151 cases, 140 entered the school and were measured on GPA. In category 3 where there were 594 applicants, 65 survived the first selection stage and were measured on the Essay, and finally 40 were offered admission. Of this latter group, only 34 were eventually measured on y. Summarizing across the three admission categories, Mathematics and Essay scores were measured on all 2483 applicants; Essay scores were measured on 474 members of this group; and GPA was measured for the 217 accepted applicants who entered and remained in the school.

In Table 2 the characteristics of the applicants who were admitted and observed on the GPA variable are presented. Thus, for all of the applicants described in Table 2, there are complete data for all variables.

INSERT TABLE 2 ABOUT HERE

An inspection of the means in Table 2 for the entrance examinations shows that different admission standards were employed for each category. It appears that the standards were most stringent for category 2, and least stringent for category 1. It should further be noted that the multiple correlations presented in Table 2 are most certainly negatively biased estimates except for category 1 where there basically are no missing data. The basis for this conclusion is the well known result, that by restricting the range of the predictor variable, the associated correlation is attenuated.

Table 3 contains the maximum likelihood estimates of the following parameters: (a) the means for the Mathematics, English, Essay, and GPA variables, (b) the multiple correlations in predicting GPA from the three predictors, and (c) the difference in the expected GPA of applicants admitted by a lottery and those admitted on the basis of the predictor variables. This latter parameter is referred to as the "expected gain from selection (EGS)." In addition, the associated standard errors of the estimates are presented. It should be noted that the parameters (means and multiple correlations) that are estimated in Table 3 apply to the entire applicant pool. For example the estimate for the mean Essay score, is an estimate of what the average Essay score within a given category would have been if all applicants were admitted. Except for category 1 (where nearly all applicants are accepted), the means in Table 3 are uniformly lower than those in Table 2 where only admitted applicants are considered.

INSERT TABLE 3 ABOUT HERE

The EGS parameter presented in Table 3 is defined in the following manner. If applicants within category j = 1,2,3 were selected by a lottery, the expected value of their average GPA would simply be the overall applicant mean or expected value for GPA, $E(GPA_j)$. This expected value can be obtained by evaluating the regression equation predicting GPA from the three predictor variables (labeled for convenience as 1,2,3) at the mean value for the predictors, i.e.,

$$E(GPA_j) = \beta_{0j} + \beta_{1j}\mu_{1j} + \beta_{2j}\mu_{2j} + \beta_{3j}\mu_{3j}. \quad (1)$$

where μ_{1j}, μ_{2j}, μ_{3j} are the means for the Mathematics, English, and Essay examinations for the population of applicants in category j. Estimates of these means are obtained from Table 3. Similarly, the expected GPA for selected applicants [$E(GPA_j|S)$] is obtained by evaluating the category regression equation at the mean predictor scores of selected applicants:

$$E(GPA_j|S) = \beta_{0j} + \beta_{1j}E(x_{1j}|S) + \beta_{2j}Ex_{2j}|S) + \beta_{3j}E(x_{3j}|S), \qquad (2)$$

where $E(x_{ij}|S)$ is the expected value of predictor i for the selected cases in category j. Estimates of these expected values can be obtained from Table 2.

The EGS parameter is then the difference in the expected values given in equations (1) and (2).

$$EGS_j = \beta_{1j}d_{1j} + \beta_{2j}d_{2j} + \beta_{3j}d_{3j} \qquad (3)$$

where $d_{ij} = \mu_{ij} - E(x_{ij}|S)$ is the difference in category j between the mean applicant score and the mean accepted applicant score for predictor i. Setting the values of the d_{ij} parameters to be the observed applicant-admitted differences (obtained from Tables 1 and 2), and replacing the population regression weights by their maximum likelihood estimates, one obtains an estimate of EGS_j, j=1,2,3. Since virtually all applicants in category 1 were admitted, the EGS parameter was estimated only for categories 2 and 3.

Two major conclusions can be drawn from the estimates in Table 3. First, the predictor variables are clearly statistically valid predictors of GPA for categories one and three. This result can be seen by testing the R^2 values for significance using a simple z test, where z = R^2/(stanc ırd error of R^2). The observed R^2 values of .34 and .40 each yield signıficant (p < .05) z values. Further evidence for predictive validity in the third category is the finding that the expected gain from selection is sigificantly different from zero (z = 7.34/2.98 = 2.46, p < .05). Second, although the results in Table 3 for the second admission category fail to reach statistical significance, they strongly suggest that the admissions variables are also valid for this category. A z test of the estimated R^2 value in category two, (z = .14/.09 = 1.56) is nearly significant (p < .06). Similarly, the estimated EGS value (1.90) for category three also approaches signifcance (z = 1.90/1.39 = 1.37, p < .08). These findings may very well be attributable to low levels of statistical power for the signifcance tests of the R^2 and EGS parameters. This issue is discussed more fully in the following section.

It is also of interest to consider the size of the standard errors and the associated confidence intervals for the estimates presented in Table 3. Although there is evidence for the predictive validıty of the predictors in all three categories, the multiple correlation and EGS parameters cannot be precisely estimated. For example, the .95 confidence intervals for the squared multiple correlations in categories 1 and 3 are [.09, .59] and [.01, .79] respectively. Similarly, the .95 interval for the EGS in category 3 is [1.50, 13.18]. The width of the confidence intervals reflects not only the large amount of missing data, but also the decision to analyze the data using the heterogeneous Σ_j model. Since far fewer parameters are estimated under the homogeneous Σ_j model, in the latter case the standard errors and the widths of the associated confidence intervals would certainly have been smaller. We will return to this issue in the final section.

All of the maximum likelihood estimates given in Table 3 were obtained under two assumptions: (a) the variance covariance matrices within each category are heterogeneous; (b) all of the missing GPA scores are MAR. We now consider the validity of each of these assumptions. The decision to present the estimates for the heterogeneous Σ_j model was based on the result of a log-likelihood ratio test. First the estimates of the category parameters (means, variances, covariances) were obtained under the assumption that the Σ_j matrices are heterogeneous. For this analysis, the EM algorithm was applied separately to the data in each category using the BMDP-AM computer program. Secondly, the parameters were reestimated under the assumption of a common variance covariance matrix. The Schlucter computer program was used for this analysis. Whereas the number of parameters estimated in the first analysis was 42 (4 means, 4 variances, and six covariances within each of three categories), the second analysis estimated only 22 parameters (4 means in each of the three categories, and the 10 elements of the common variance covariance matrix). The null hypothesis of a common Σ matrix was tested by computing a chi square log-likelihood ratio test with 42-22=20 degrees of freedom. This statistic was computed by taking the log of the ratio of the likelihood of the data given the two sets of estimates. The resulting chi square was highly significant leading to the rejection of the equal Σ_j model.

Consider the second assumption that all of the missing GPA scores are MAR. As previously noted, this assumption is tenable for all but the 18 applicants who although offered admission, either declined or left the school prior to the time that the GPA variable was measured. To investigate the possible effect of violations in the MAR assumption, a sensitivity analysis was performed where different values were imputed for these missing scores and the parameter values reestimated. When the MAR assumption holds, the regression equations predicting GPA (obtained from the complete data of the admitted applicants within each category) will provide unbiased estimates of the GPA scores of the 18 missing cases. However, when the MAR assumption does not hold, i.e. when these GPA scores are not missing simply as a function of the predictor variables, the previous equations will yield biased predictions (Gross, 1987). To allow for the possibility that these 18 scores are not MAR, we imputed the missing values using the category regression equation together with an adjustment factor. More specifically, given an applicant from category j, with predictor scores x_1, x_2 x_3, the predicted GPA was obtained using the complete case regression equation. This predicted value was then modified by adding an adjustment factor. This factor was either positive or negative one residual standard deviation unit. In other words, the 18 missing values were estimated to be either systematically higher or lower than their predicted values. The squared multiple correlations and the EGS estimates for the three categories were then recomputed using the data set which included the imputed values. The results of these computations presented in Table 4 suggest that the analysis is fairly robust with respect to the assumption that the 18 missing GPA scores are MAR. The estimates of the squared multiple correlations and EGS values are substantially unchanged when different values (one residual standard deviation unit above and below the predicted value) are imputed for these missing scores. For example, in category 3, the squared multiple correlation varies from the original maximum likelihood estimate of .40 by no more than .05 units under the two imputations. Similarly, the EGS values are changed by less than one point.

4. Conclusions

In analyzing any data set which contains missing data one must introduce assumptions concerning the missing data process to obtain statistically accurate estimates of the underlying parameters. The sensitivity of the analysis to these assumptions is clearly a function of the proportion of missing data. Violations of the underlying assumptions may be only a minor problem when there are relatively little missing data, but can lead to highly biased estimates when the proportion of missing data is high. In investigating the predictive validity of a test battery for a highly selective institution, there will be by definition large amounts of missing data. The key assumption in this type of analysis is that the missing data can be accounted for in terms of the observed and measured variables, i.e., the missing data are missing at random. The proposed model represents an attempt to assure that this assumption will be satisfied by measuring not only the predictor variables, but also noting the admission category of the applicant. While the predictors alone cannot always account for the missing data, the measurement of the predictor variables together with the admission category may very well yield a data set where the missing data are missing at random.

Although the proposed model is quite general and can be employed when there are missing data on the predictors, criterion and the admission category, the model is still potentially limited in two ways. As previously noted, the missing criterion scores of those applicants who decline an admissions offer or soon drop out, may not be MAR. In other words, there may be additional variables (statistically related to the criterion) which are responsible for these missing scores. For this problem, we have suggested that the investigator perform a sensitivity analysis where a range of reasonable values for these missing criterion scores are imputed and the parameters reestimated. For the data set considered in the present paper, the sensitivity analysis suggested that the model was robust with respect to the assumption that the missing criterion scores of admitted applicants were MAR.

The second limitation of the proposed model is that although it yields unbiased estimates of the relevant parameters, both the precision of these estimates as well as the power of any significance tests based on these estimates may not be high due to the large amount of missing data as well as the form of the missing data patterns. The problem of precision was noted in terms of the rather wide confidence intervals obtained for the squared multiple correlations and the EGS parameter. The issue of low power is most clearly seen in the results for the second admission category where the results only approached significance. It can be argued that this finding is attributable to low levels of statistical power. More specifically, for the second admission category, complete data could be observed for only 140 of 1845 applicants. Further, within this complete case group, there was considerable restriction in the range of the predictor variables. The ratios of the standard deviations of each predictor variable in the selected group to the corresponding standard deviation for the applicant group were .52, .67, and .70 for the

Mathematics, English, and Essay examinations respectively. Although the complete sample size of 140 is not small, the high levels of range restriction within this sample can easily result in low power.

The problems of wide confidence intervals and low power in testing hypotheses are most likely to occur when the heterogeneous Σ_j model is employed. In this case, the parameter estimates are obtained using only the data from a single admission category at a time. It is clear however, that the standard errors would be much smaller if the pooled or homogeneous Σ_j analysis were employed. Thus, the proposed model should be most useful when the data are consistent with the hypothesis of common variance covariance matrices across the admission categories. In addition to the choice of the heterogeneous or homogeneous Σ_j model, the precision of the estimates will be a function of factors such as the proportion of missing data, the number of admission categories, the overall sample sizes for each admission category, and the form of the missing data patterns, i.e. the degree of range restricition. It would be of practical value to provide some general guidelines in terms of these factors for identifying the data sets where the model can be expected to provide reasonably precise estimates. The construction of these guidelines would clearly be a useful area for future research.

References

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

BMDP Statistical Software Manual. (1990). University of California Press: Berkeley.

Gross, A.L. & McGanney, M. (1987). The restriction of range problem and nonignorable selection processes. Journal of Applied Psychology , 72, 604-610.

Heckman, J.J. (1976). The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a simple estimator for such models. Annals of Economic and Social Measurement, 5, 475-492.

Heckman, J.J. (1979). Sample selection bias as a specification error. Econometrika, 47, 153-161.

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

Little, R.J.A, & Rubin, D.B. (1987). Statistical analysis with missing data. New York: John Wiley.

Mood, A.M., Graybill, F.A., & Boes, D.C. (1974) Introduction to the theory of statistics. New York: McGraw-Hill.

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

Rubin, D.B. (1976). Inference and missing data. Biometrika, 63, 581-592.

Table 1

Observed Characteristics of Applicants

Categ.	N_{applic}[a]	N_{offer}[b]	N_{attend}[c]	Mean Math	Mean English	Mean Essay
1	44	44	43	51.14 [13.06] (44)	42.91 [8.14] (44)	84.52 [31.29] (44)
2	1845	151	140	39.42 [12.21] (1845)	37.42 [8.08] (1845)	91.45 [30.52] (365)
3	594	40	34	34.27 [12.01] (594)	33.85 [7.53] (594)	89.37 [32.59] (65)

Note. Standard deviations are in brackets, sample sizes are parenthesized.

[a] N_{applic} = number of applicants.

[b] N_{offer} = number of applicants offered admission.

[c] N_{attend} = number of applicants offered admission who attended and were measured on y.

Table 2

Observed Characteristics of Admitted Applicants

Categ.	Mean Math	Mean English	Mean Essay	Mean GPA	R^2 [a]
1 (N=43)	50.67 [12.84]	42.88 [8.25]	85.05 [31.47]	86.50 [4.76]	.32
2 (N=140)	59.54 [6.36]	49.11 [5.41]	102.17 [21.41]	89.72 [4.12]	.06
3 (N=34)	55.74 [6.90]	45.23 [6.92]	103.14 [25.52]	88.14 [4.98]	.15

Note. Standard deviations are in brackets.

[a]R^2 The squared multiple correlation predicting GPA from Math, English, and Essay for applicants measured on all variables.

Table 3

Maximum Likelihood Estimates and Standard Errors

Category	Mean Math	Mean English	Mean Essay	Mean GPA	EGS [a]	R^2 [b]
1	51.14 (2.36)	42.91 (1.34)	84.52 (5.22)	86.62 (0.75)		.34 (.13)
2	39.42 (0.29)	37.42 (0.19)	85.87 (4.96)	87.82 (1.44)	1.90 (1.39)	. 14 (.09)
3	34.27 (0.53)	33.85 (0.33)	58.01 (14.19)	80.80 (3.16)	7.34 (2.98)	.40 (.20)

Note: Standard errors are in parentheses.

[a] EGS The difference in the expected value of GPA between applicants selected in terms of the x variables and those selected by a lottery.

[b] R^2 The maximum likelihood estimate of the squared multiple correlation in predicting GPA from Math, English, and Essay.

Table 4

Sensitivity Analysis

Category	R^2_{MAR} [a]	EGS_{MAR} [b]	R^2_A [c]	R^2_B [d]	EGS_A [e]	EGS_B [f]
1	.34		.34	.31		
2	.14	1.90	.12	.18	1.31	2.50
3	.40	7.34	.37	.45	6.73	7.96

[a] R^2_{MAR} Squared multiple correlation value assuming the "18 cases" are MAR.

[b] EGS_{MAR} Expected gain from selection assuming the "18 cases" are MAR.

[c] R^2_A Squared multiple correlation value where the 18 missing GPA's are imputed to be one residual standard deviation unit above the predicted value.

[d] R^2_B Squared multiple correlation value where the 18 missing GPA's are imputed to be one residual standard deviation unit below the predicted value.

[e] EGS_A Expected gain from selection where the 18 missing GPA's are imputed to be one residual standard deviation unit above the predicted value.

[f] EGS_B Expected gain from selection where the 18 missing GPA's are imputed to be one residual standard deviation unit below the predicted value.