Order Number 9224843

An application of the EM algorithm in analyzing the CUNY open-admissions study missing-data

Na, Hazon, Ph.D.

City University of New York, 1992

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

AN APPLICATION OF THE EM ALGORITHM IN ANALYZING THE CUNY OPEN-ADMISSIONS STUDY MISSING-DATA

by

HAZON NA

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.

1992

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

4/23/92
Date

Chair of Examining Committee

4/29/92
Date

Executive Officer

Dr. Alan L. Gross, Chairman

Dr. David E. Lavin

Dr. Roger Millsap

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

AN APPLICATION OF THE EM ALGORITHM IN ANALYZING
THE CUNY OPEN-ADMISSIONS STUDY MISSING-DATA

by

HAZON NA

Advisor: Prof. Alan L. Gross

The present study is based on an analysis of a sample from the CUNY open-admissions data set. The data set consisted of two portions, an original sample and a follow-up sample which contained only 14% of the original cases. Not only were data missing for those cases not in the follow-up sample, but the original sample variables were not completely observed. The data set is basically multivariate with both incomplete continuous and categorical variables. In analyzing such a data set, many researchers typically use ad hoc approaches that lack theoretical bases. For example, deletion or substitution methods are offered as a routine treatment for missing values before performing an analysis in many statistical packages.

It is important to note that deletion methods using only respondents with no missing values may yield biased results, unless the complete cases can be viewed as a completely random subsample of the original sample observations. A more realistic approach is to assume that the missing data are not missing in a completely random fashion, but rather are missing at random as a function of known subject characteristics. Further, given this more realistic assumption concerning the missing data process, one could apply Maximum Likelihood methods to estimate the parameters of interest. The Maximum Likelihood method was used in the present study.

In this study, the Maximum Likelihood estimates for means, variances, and correlations were obtained by implementing the Estimation-Maximization (EM) algorithm suggested by Little & Schulucter (1985). These Maximum Likelihood estimates were compared with the estimates obtained from three different ad hoc methods; Pairwise deletion, Listwise deletion, and Weighting analyses.

Although the results show some differences in terms of correlation estimates, there was little evidence that the methods yield different estimates of proportions, means and standard deviations. Possible explanations for this result are discussed. In general, however, the ad hoc and Maximum Likelihood methods will not agree.

ACKNOWLEDGEMENTS

I wish to thank many individuals and institutions for giving me the part-time jobs to support school and living expenses since I came to New York. I am truly grateful to people who cared about me and have taken me in their trust.

First I would like to give an acknowledgement to the City University of New York Graduate School and University Center for granting me the admission and providing me with financial aids. On campus, thanks are due to the faculty, the staff, and the colleagues of the Educational Psychology department, the staff of the Computer Center and the faculty and the staff of the Mina Rees library, especially, the acquisition department. The Baruch College Educational Computer Center too has given me the secure employment.

I was also benefitted by the generosity of the Korean community in the New York City, and particularly blessed with the presence of the Korean Church in the community and the Christian fellowship.

CONTENTS

Chapter I

INTRODUCTION

Twenty years ago, the City University of New York (CUNY) adopted a policy which guaranteed admission to every graduate of the city's high schools. The goal of this program was to offer college degree and to provide necessary remedial education for underprepared students. At the outset, CUNY's open-admissions policy was criticized as a threat to academic standards and an unlikely remedy for socioeconomic immobility. Lavin et al. (1981) extensively examined the effects of the open-admissions policy on the university's ethnic composition on placement (in two- or four-year colleges) and on the academic performance of different groups.

The analyses were based on data collected on samples of CUNY students prior to their college admission and during their college years. The data set included measures of social and educational as well as collegiate achievement. Basing their conclusions on a detailed study of approximately 35,000 students among the first three classes (a total population of almost 100,000 students) who were followed from Fall 1970 through Spring 1975, the authors

argue instead that there has been no definitive evidence of a decline in standards and that the policy has been successful in reducing educational inequality.

The ultimate question to be answered now is whether the students who entered CUNY during the period of the open-admissions policy benefitted from the experience. More specifically, what have been their highest levels of educational achievement; how have they managed in the labor market; has their college experience made an impact on the socioeconomic mobility of disadvantaged groups? In 1984, approximately 14 years after their initial entry into CUNY, Lavin and his colleagues embarked on a follow-up study to answer these questions. As a result, a longitudinal data set was generated containing the students' educational attainments and work experiences over a period from the late 1960s (when they typically were attending high school) through the early 1980s.

The focus of Lavin et al.'s follow-up study was the sample of nearly 35,000 students among the 1970-1972 CUNY entering freshmen for whom the data were collected on various sociodemographic and academic variables in the original study. They are labeled as the "original sample" in Lavin et al.'s study. For his follow-up study, Lavin made an effort to survey those 35,000 students and obtain additional data with respect to various educational attainments, life and work variables. However, a little less than 5,000 of the original sample responded to a follow-up questionnaire. Thus there was an extensive

missing data problem. The data set is multivariate with both incomplete continuous and categorical variables. Even for the 5,000 students reached in the follow-up study, there are missing data values. Further the unreached 30,000 students obviously have no follow-up data and are also missing some of the original data.

The proposed study will address the issue of multivariate incomplete data like the CUNY open-admissions data set where some of variables are continuous and some are categorical. Lavin et al.'s study used only the 5,000 participants of the follow-up study, discarding from the analyses those who did not respond. His method simply resorted to weighting the observed follow-up data in order to obtain a reasonable representation of the original sample. The goal of his study was to estimate means in assessing education and work outcomes. It is important to note, however, that analyses using only complete cases may yield biased results, unless the complete cases can be viewed as a completely random subsample of the original sample observations. A more realistic approach is to assume that the missing data are not missing in a completely random fashion, but rather are missing at random as a function of known subject characteristics.

The purpose of the present study is to analyze the data set under the more realistic assumption that the missing data are not missing completely at random. In this study, the Maximum Likelihood estimates of means, variances, and correlations will be obtained by implementing the

Estimation-Maximization (EM) algorithm suggested by Little & Schulucter (1985). The EM algorithm is an iterative procedure that is general enough to be used with any pattern of missing data to obtain Maximum Likelihood estimates and is a more formal solution than the weighting procedure that is relatively unsystematic.

In the following chapter II, the structure of the CUNY open-admissions data set is described in detail. At the same time, an illustration is given of the weighting procedure carried out in the Lavin et al.'s study. Various methods of handling missing data are reviewed from the literature and the Maximum Likelihood approach by means of EM algorithm is proposed in chapter III. In chapter IV, the EM algorithm is presented in terms of a simple example for missing data with mixed continuous and categorical variables. Chapter V describes the specific parameters which will be estimated from the CUNY open-admissions data set. The results from various types of missing data analyses are presented in chapter VI. Finally, chapter VII contains the summary and disscussion.

Chapter II

THE CUNY OPEN-ADMISSIONS STUDY DATA SET

The data set to be studied in the present study was obtained as a sample from N = 102,780 students who entered CUNY's 17 senior and community colleges in 1970, 1971 and 1972, just after the open admissions policy began. The original sample consisted of N_1 = 34,731 cases. In the 1984 follow-up study, an attempt was made to obtain additional data on all of these cases. Only N_2 = 4,988 of the original N_1 = 34,731 could be obtained. The descriptions of the variables in both data sets are summarized from Lavin's follow-up study.

The Original Data (N_1 = 34,731)

Most individuals in the original sample have records on two major sets of variables. The first set of variables consists of freshmen surveys conducted when the three freshmen classes entered CUNY. The questionnaires were administered to the 1970, 1971, and 1972 freshmen classes

asking for a wide range of information about students' sociodemographic characteristics, attitudes, and aspirations. The second set of variables contains official academic records of these participants of the freshmen surveys. These records include information on (i) high school background, (ii) college admissions, and (iii) college academic achievement covering the period from the fall 1970 through 1975.

Set 1: Freshmen Surveys

Two major types of variables were included in the surveys. The first pertain to social origins and demographic characteristics, while the second refer to aspirations and attitudes. The social origins and demographic variables are as follows:

1. Age at entry to CUNY
2. Gender
3. Family income, ranging from less than $4,000 to $15,000 or more.
4. Educational attainment of mother and father, ranging from less than grammar school through a postgraduate degree.
5. Ethnicity. Three ethnic categories are mutually exclusive: black, Hispanic, and Non-hispanic white.

These are the variables pertaining to aspirations and attitudes:

1. Degree aspirations. Students indicated the highest degree to which they aspired: no degree, associate's degree, a bachelor's degree, master's degree, a Ph.D., M.D., LL.B., or other professional degree.
2. Academic self-rating. Students were asked to rate their academic ability relative to other students in their college class.

3. Employment while in college. Students indicated whether they were or were not working at an outside job while going to college.
4. Job plans after college. Respondents were asked whether they had decided on a career after they finished college.
5. Reasons for going to college. Students rated the importance of various reasons for attending college.

Set 2: Academic Records

(i) High School Background

High school transcript information collected by the University's centralized admissions office contains information on numerous variables; three are described here as major indicators of secondary school background.

a. High school average. This is a measure of the student's grades in college preparatory courses (e.g., English, mathematics, science). This variable provides the basis for distinguishing those whose entry to the University was made possible by open admissions from those who would have qualified even without the policy.

b. High school rank. This variable indicates the student's standing relative to other students in the same high school graduating class. Along with high school average, this criterion determined eligibility for a senior college under the open-admissions program.

c. Total number of units in college preparatory courses taken. They are the courses considered by CUNY and most universities to form the core of college preparatory work. Illustrative are courses in English, mathematics, sciences, foreign languages, history, and social studies.

d. The Open Admissions Test. To obtain an overview of entering students, academic preparedness, and thus an indication of the need for remedial services, CUNY administered university-wide tests of academic skills to graduating high school seniors in 1970 and 1971 (1972 seniors did not take these tests). Two tests were used. One measured reading comprehension; the other assessed numerical skills.

(ii) College Admissions Data

At the onset of the open-admissions policy, the CUNY system constituted 17 undergraduate colleges. Eight were four-year senior colleges; the remainder were two-year community colleges. This data show the level of the college in which each student wanted and was placed on entry.

(iii) Collegiate Academic Achievement

After each semester, including summer sessions, each CUNY campus transmitted to the University's central office computer tapes containing information on the courses students took, their grades, and the number of credits each course carried. Each student's academic performance was constructed for the period 1970-1975. The following variables are included:

a. Credits and grades. The file shows the total number of credits a student registered for at CUNY as well the total earned. The cumulative grade point average for the entire period of the student's enrollment is included. The file also shows the grade point average earned by students in each year at CUNY.

b. Remediation data. CUNY's open-admissions policy was accompanied by a large remedial and compensatory effort designed to upgrade the academic skills of the many underprepared students who entered after the program began. For this study the remediation variables include the number of remedial courses taken, and the number passed.

c. Enrollment and graduation data. The file indicates enrollment status (registered, not registered) during each semester over the period from fall 1970 through spring 1975. It also indicates whether students had graduated by June 1975 and the degree(s) received.

The Follow-Up Data (N_2 = 4,988)

The third data set is the 1984 follow-up survey that provides the additional data for the original sample. The questionnaire asked for information about former students'

lives, including ultimate educational attainments (at CUNY and elsewhere), labor market experiences as of 1984, and attitudes toward and satisfaction with various aspects of their life situations. These data from the follow-up survey were integrated with the original data so the record of each respondent (N_2 = 4,988) also contains information on high school background, academic performance in CUNY as well as data on social origins and attitudes. Thus, in the merged data set, only N_2 = 4,988 out of N_1 = 34,731 have the follow-up data in addition to the original data and $N_1 - N_2$ = 29,743 have missing values on all of the follow-up variables. Further both the original data and the follow-up data were incomplete, having their own missing values.

The educational attainment items:

1. Degrees earned, ranging from a high school diploma through advanced and professional degrees.
2. Further education beyond most recent degree.

Data on labor market experience and work situations:

1. Employment status while an undergraduate.
2. Year of first full-time job.
3. Number of years employed since 1970.
4. Employment status in 1978 and 1984.
5. Job title in 1978 and 1984.
6. Salary for job held in 1978 and 1984.
7. Type of organization (e.g., bank, hospital, self-employed, etc.) worked for in 1978 and 1984.
8. Quality of work. The questionnaire asked for information about several characteristics of respondents' current jobs.
9. Job source for current job and job held in 1978. Respondents indicated the most important source (friend, relative, employment agency, etc.) in helping them to get a job.
10. Work layoffs and unemployment benefits.
11. Child care and work.
12. Job discrimination.

Attitudes toward work and other aspects of individuals' life:

1. Marital status, and if married, the employment status, job title, and salary of the spouse.
2. Number of dependent children living with the respondent.
3. Attitude toward work. Respondents were asked how they felt about a number of aspects of work.
4. Self-image and attitudes. A number of items were designed to tap respondents' feelings about themselves.
5. Behavioral dimensions. A number of items asked respondents to describe their civic involvement, their cultural involvement, and personal change.
6. Life satisfaction. Respondents were asked to indicate their satisfaction with numerous aspects of their life situations.

Previous Analysis of the Open-Admissions Data Set

Using this merged data set, Lavin focused on the follow-up sample of N_2 = 4,988, discarding from the analysis those who did not respond to the follow-up study. An attempt was made to generalize the results to the total population (N = 102,780). However, it was clear that the N_2 = 4,988 respondents were not a random representative sample of the data base from the original sample (N_1 = 34,731). The follow-up study respondents appeared to be more likely to graduate and contained more women and a greater percentage of younger students than the original sample. In order to adjust for these differences, Lavin weighted the follow-up

sample so that it closely approximated the characteristics of the original sample on variables common to both N_1 and N_2. The decision to compare the follow-up sample with the original sample rather than with the population of N = 102,780 is based on the earlier study (Lavin, Alba, and Silverstein, 1981) which indicated a close correspondence between the original sample and the population. Although, the original sample contained greater proportions of students with strong high school records and college academic performance as compared with the total population, these differences were slight.

The weighting procedure in Lavin's study was carried out as suggested by Berk (1983). The procedure involves predicting the probability (P) that any given individual in the original sample (N_1 = 34,731) would have responded to the follow-up study (S) based on the known characteristics $(X_1\ X_2\ \ldots\ X_i)$ of that individual in the original data (X_i). The variables in the original data that were found to contribute to the probability of falling into the follow-up sample are as follows: Race/Ethnicity, Age, Gender, High School Average, Income on Entry to CUNY, Year of Entry to CUNY, Level of Entry to CUNY, Average Credits Earned per semester while attending CUNY, and Graduation Status which was updated through 1982 for the weighting procedure. Continuous variables, such as income, were categorized and each category was made into a dichotomous variable to be included in the analysis. A missing data

category for these variables was represented by means of a dummy variable. For example, 15% of the original sample who have no income data were treated as a separate income group.

The set of known $(X_1\ X_2\ \ldots\ X_i)$ characteristics were used in the construction of a logistic multiple regression model:

$$P(S \mid X_1 X_2 \ldots X_i) = \exp(\Sigma X_i \beta_i) / [1 + \exp(\Sigma X_i \beta_i)] . \qquad (1)$$

The model computes the probability that S = 1 for each individual in the follow-up sample (N_2 = 4,988). The inverse of this probability was then multiplied with the follow-up variables (Y_j) for each subject:

$$Y^*_j = Y_j / P(S \mid X_1\ X_2\ \ldots\ X_i) . \qquad (2)$$

The procedure adjusts the distribution of the variables in the follow-up sample. The adjusted follow-up variables $(Y^*_1 Y^*_2 \ldots Y^*_j)$ and the original data $(X_1\ X_2\ \ldots\ X_i)$ for those who responded to the follow-up survey (S = 1) were the basis of Lavin's follow-up study. Those who did not respond (S = 0) were discarded from the study.

To illustrate the weighting procedure, take a simple univariate example and suppose we are interested in estimating the mean (μ) of income for a population of 10. We assume that only a sample of 3 have responded to our survey. Since we have missing data on the rest of seven cases, we can not generalize the mean income computed from the sample to the population. It is necessary to assess the representativeness of our sample. Suppose we have

information on the gender of our population; 1 male and 2 females have responded to our survey that was sent to the total population of 4 males and 6 females. Now we can compute that the probability a male would respond to the survey is 1/4 and for a female, 1/3. Consider that our male respondent has reported his income as $40,000 and the two females have reported $25,000 and $35,000 respectively. Rather than simply summing the respondents' income and then dividing the sum by the number of respondents, we weight each individual's income multiplying by the inverse of their probability of being in the sample. We then add up these values and divide by the sum of the reciprocal of the weights:

$$\mu = \frac{40{,}000 \times \frac{1}{4}^{(-1)} \times 25{,}000 \times \frac{1}{3}^{(-1)} \times 35{,}000 \times \frac{1}{3}^{(-1)}}{1 \times \frac{1}{4}^{(-1)} + 2 \times \frac{1}{3}^{(-1)}}. \qquad (3)$$

Thus the weighted income of our representative sample would be $34,000 rather than $33,333 which is the average income of three respondents.

When both distributions are compared, the weighted follow-up sample shows a reasonable representation of the original sample as far as the original variables are concerned. The effects of weights on the follow-up variables are not clear, since the values for all follow-up variables are missing for the original sample who did not respond to the follow-up study. Since the weights were created based on the variables recorded for the whole

original sample, the analysis based solely on the follow-up sample may bias the outcomes of the research. It is more reasonable to utilize formal statistical methods (i.e. Maximum Likelihood approach) to analyze the data set with missing values rather than relying on the ad hoc methods such as weighting procedures to make the data set "look" comparable.

Chapter III

REVIEW OF THE LITERATURE AND PROPOSED METHOD

Basically three types of missing-data pattern have been studied in the literature. Most studies have illustrated the patterns in the context of continuous variables and with two or three variables (Anderson, 1957; Beale, 1970; Buck, 1960; Lord, 1955; Matthai, 1951; Wilks, 1932). Multivariate extensions are also explored (Anderson, 1957; Beale & Little, 1974; Bhargava, 1962; Dempster, 1969; Glasser, 1964; Haitovsky, 1968; Marini, Olsen, & Rubin, 1985; Rubin, 1974; Rubin & Thayer, 1978; Trawinski & Bargmann, 1964). In some studies the variables are categorical (Chen & Fienberg, 1974; Fuchs, 1982) or mixed, with both categorical and continuous (Krzanowski, 1980; 1982; Little & Schulcter, 1983).

Type I : Data Missing On Only One Variable

The simplest pattern is where data are missing on only one variable, and all other variables are completely observed. In the case of two variables (V_1, V_2) , the missing-data pattern is illustrated in Figure 1. The data consist of n_1 observations on both variables, V_1 and V_2, and

$N - n_1$ observations on V_1 alone. Thus the values from V_{2n1+1} to V_{2N} are considered as missing data. The binary response indicator (R) for the variable (V_2) takes the value 1 if it is recorded and the value 0 otherwise. This type of incomplete data assuming an underlying bivariate normal distribution was first considered by Wilks (1932). The pattern has been estimated with ML methods by Mattahi (1951), Lord (1957), and Anderson (1957). Another area that addresses this type of data is the restriction of range problem (Brewer & Hills, 1969; Greener & Osburn, 1979; Gross, 1982).

V_1	V_2	R
v_{11}	v_{21}	1
v_{12}	v_{22}	1
.	.	.
.	.	.
.	.	.
.	v_{2n1}	1
.	?	0
.	?	0
.	.	.
v_{1N}	?	0

Figure 1. Two variables with missing data in one

Type II : Data Missing On More Than 1 Variable With One Observed More Than Another

If data are missing on more than one variable as is likely in a multivariate data set, data can sometimes be arranged hierarchically from fully observed to less observed variables (e.g. Marini, Olsen and Rubin, 1980). Figure 2 illustrates such pattern for 3 variables with V_2 less observed than V_1 which is fully observed, and V_3 less

observed than V_2. Response indicators are for the V_2 and V_3 variables. In the literature, this pattern has been labeled as "nested" (Hartley and Hocking, 1971) or "monotonic" (Bhargava, 1962). The monotonic pattern has unique implications in ML estimation since the likelihood function for this pattern factorizes into separate terms containing distinct parameters. Since the parameters are distinct, each terms can be separately maximized. The trivariate case was studied by Matthai (1951) and Lord (1955). The multivariate generalization for the normal distribution was described by Anderson (1957), Bhargava (1962), and Rubin (1974).

V_1	V_2	V_3	R_2	R_3
v_{11}	v_{21}	v_{31}	1	1
v_{12}	v_{22}	v_{32}	1	1
.	.	.	.	.
.	.	.	.	.
.	.	v_{3m1}	.	1
.	.	?	.	0
.	.	?	.	0
.	.	.	.	.
.	v_{2m2}	.	1	.
.	?	.	0	.
.	?	.	0	.
.	.	.	.	.
v_{1N}	?	?	0	0

Figure 2. Monotone Missing-Data Pattern with Three Variables

Type III : Data Missing On More Than 1 Variable With Non-monotonic Pattern

In most cases, a non-monotonic pattern of missing data will occur. Such a pattern is shown in Figure 3. The incomplete bivariate data contain missing values on both

variables and can be partitioned into three groups. The first group has observations on variables, V_1 and V_2, the second group, V_2 and V_3, the third group, V_1 and V_3. Unlike the monotonic pattern, the likelihood function for the data in each group is complicated since the parameters underlying the data are not distinct. As a result, the full likelihood for the data in all three groups has to be maximized simultaneously. For the multivariate data, some studies demonstrate iterative methods of obtaining ML estimates (Beale & Little, 1974; Dempster, 1969; Glasser, 1964; Haitovsky, 1968; Trawinski & Bargmann, 1964). Afifi & Elashoff (1967, 1969) surveyed the methods of ML estimation with the normal distribution in the literature.

V_1	V_2	V_3	R_1	R_2	R_3
v_{11}	v_{21}	?	1	1	0
.	.	.	.	.	.
v_{1n1}	.	?	1	.	0
?	.	v_{3n1+1}	0	.	1
?	.	v_{3n1+2}	0	.	1
.	.	.	.	.	.
?	v_{2n2}	.	0	1	.
v_{1n2+1}	?	.	1	0	.
v_{1n2+2}	?	.	1	0	.
.	.	.	.	.	.
.	.	.	.	.	.
v_{1N}	?	v_{3N}	1	0	1

Figure 3. Non-Monotonic Pattern with Three Variables

The literature dealing with the analysis on these three missing data pattern can be categorized into two approaches; ad hoc procedures, and methods based on formal statistical theory. The latter approach is typified by the

work of Little & Rubin (1987). In the next section we review the ad hoc procedures. The following section describes the formal methods.

Ad Hoc Procedures of Analyzing Incomplete Data

These are straightforward methods that have been practiced without a theoretical basis, and thus tend to be ad hoc. By and large, these methods "maneuver" missing values in order to prepare the data in a way that the standard statistical analysis for complete data can be applied. They are widely used among researchers and data analysts partly because the procedures are easy to handle, and also are available in statistical computing softwares, such as SPSS-X (1988). For example, deletion or substitution strategies are offered as a routine treatment for missing values before performing an analysis in SPSS-X (1988).

The deletion strategy is a complete-data method that deletes all the cases with missing data in any variable and analyze only the cases with complete data across variables, which inevitably reduces sample size. It is termed "listwise" deletion (Kim & Curry, 1977; Norusis, 1983), or "complete-case" analysis (Little & Rubin, 1987). In the three types of missing-data pattern illustrated above,

listwise missing-data treatment will delete any cases where the response indicator (R) takes the value 0 on any variable. If the data are complex with non-monotonic missing data pattern (e.g. Type III), there will be greater loss of data by elimination. The deleted cases may also significantly differ from the remaining ones.

A variation to the complete-data method is to use the cases that are observed at least on the pairs of variables under the study. This is called "pairwise" deletion (Kim & Curry, 1977; Norusis, 1983), or "pairwise available-case" method (Little & Rubin, 1987). Pairwise deletion usually brings changes in the sample base from one pair of variables to another. For example, in Figure 3, the number of cases changes from n_1 (for the pair of V_1 and V_2) to $n_2 - n_1$ (for the pair of V_2 and V_3), or to $N - n_2$ (for the pair of V_1 and V_3). In incomplete multivariate data where there are more than 3 variables, anomalous relationships between coefficients may occur since different cases will be used to estimate measures of covariation (Little & Rubin, 1987; Norusis, 1983). The lists of problems and limits inherent in the deletion strategy are documented by Haitovsky (1968), Kim & Curry (1977), and Marini, Olsen, & Rubin (1974).

Mean imputation is another option offered in SPSS-X for the treatment of missing-data. It substitutes missing values with either the total sample mean or subgroup mean and the outcome is considered as complete data. Mean imputation is one of many individual imputation methods which are used extensively for survey data. Other

imputation methods include hot deck imputation, where the missing values are replaced by selected observed values; regression imputation, in which missing values are predicted by regressing the variables that contain missing values on the known variables fully observed; and multiple imputation methods which replace each missing values with two or more imputed values.

Although imputation is a versatile means of handling missing data problems, Sedransk (1985) points out that "no standards exist for deciding when the use is appropriate, what is the optimal method, or what is the maximum level of nonresponse that can be dealt with in this way". Dempster & Rubin (1983) argue it is dangerous believing that the imputed data are complete after all, and because it can not distinguish situations where the problems resulting from imputation are minor or when there exits substantial bias. Kalton & Kasprzyk (1982) and Bailar & Bailar (1983) also reviewed various imputation methods in survey research. Little & Rubin (1987) discussed these methods from the missing-data perspective.

Weighting procedures are usually adopted when one or more cases have completely missing values across all variables or a set of variables under study in survey data (Kalton & Kasprzyk, 1982). The procedure involves producing the weights which are inversely proportional to the probability of responding cases in a subset of variables. Previous analyses of the CUNY open-admissions data set have treated the missing-data problem with weighting procedures.

Weights were produced based on all cases in the data set. Parameter estimation for variables in the follow-up sample was carried out by weighting the scores of the respondents in the follow-up sample.

The problems with ad hoc procedures in analyzing the data with missing value are clear in that: there is a great loss of information; the estimates are inconsistent; they often violate assumptions and require ad hoc adjustments to generate satisfactory estimates; and it is not easy to distinguish situations when the methods work from situations when they fail (Little & Rubin, 1987). An ideal approach to the missing-data treatment would be specifying a statistical model for the incomplete data. Inferences are based on the likelihood under that model and the missing-data mechanism. Population parameters are estimated by procedures such as Maximum Likelihood (Little & Rubin, 1987).

The drawbacks of this model-based approach is that except for very simple situations it often depends on iterative numerical methods in order to obtain estimates, and model specifications are complex for general patterns of missing data. But the advantages of this approach are: it can be applied to both simple and complex situations; it avoids ad hoc aspects of quick methods by specifying a statistical model; it uses a greater amount of available information; large sample estimates of standard error are available with the missing values taken into consideration (Kim & Curry, 1977; Little & Rubin, 1987). We now consider these more formal approaches.

Rubin's Model

Rubin (1976) has argued that a systematic approach to missing data analysis requires explicit assumptions concerning the mechanism that led the values to be missing. He characterized the missing-data mechanism in terms of a random variable (R) that indicates whether a variable is observed (R = 1) or missing (R = 0) for each unit of analysis. Suppose we choose (V) to stand for the values of the variables in an incomplete data set. If V_{obs} are the values for observed data, and V_{miss} are the potential values for missing data, the response mechanism is defined in terms of conditional probability distribution of a response indicator (R) given both observed data (V_{obs}) and missing data (V_{miss}) : $P(R \mid V_{obs}, V_{miss})$. Then a missing-data set can be characterized as being actually comprised of $V = (V_{obs}, V_{miss})$ and (R).

Rubin's model formulates a missing-data set in terms of a joint probability distribution of the data (V) and a response indicator (R). The conditional distribution of (R) given the data (V) is indexed by an unknown parameter ϕ, and the parameters underlying the data (V) are denoted by θ. Then the joint distribution is specified as:

$$p(V, R \mid \theta, \phi) = p(V \mid \theta)\, p(R \mid V, \phi). \qquad (1)$$

In order to estimate the parameters for the actual observed

data, the marginal distribution of the observed data is obtained by integrating the joint density of (V, R) over the missing values (V_{miss}). Thus the loglikelihood (L_{obs}) of parameters (θ, ϕ) underlying the observed data is described as:

$$\begin{aligned} L_{obs}(\theta,\phi) &= \log P(V_{obs}, R \mid \theta, \phi) \\ &= \log \int_{V_{miss}} P(V_{obs}, V_{miss}, R \mid \theta, \phi) \\ &= \log \int_{V_{miss}} P(V_{obs}, V_{miss} \mid \theta)\, P(R \mid V_{obs}, V_{miss}, \phi) . \end{aligned} \quad (2)$$

Rubin emphasizes that the appropriate procedure for obtaining Maximum Likelihood (ML) estimates when missing data are present depends upon what assumptions can be made concerning the missing-data mechanism, the second component of the above equation (2), $P(R \mid V_{obs}, V_{miss}, \phi)$. It is important to note that, given the observed data (V_{obs}), if the distribution of missing-data mechanism (R) does not depend on missing values (V_{miss}), that is,

$$P(R \mid V_{obs}, V_{miss}, \phi) = P(R \mid V_{obs}, \phi), \quad (3)$$

then the missing data are considered missing at random (MAR) and the loglikelihood equation (4) is the same as:

$$\begin{aligned} L_{obs}(\theta,\phi) &= \log p(V_{obs}, R \mid \theta, \phi) \\ &= \log P(V_{obs} \mid \theta) + \log P(R \mid V_{obs}, \phi) \\ &= L_{obs}(\theta) + L_{obs}(\phi) . \end{aligned} \quad (4)$$

Given the MAR assumption, and in addition that the parameters in the equation (2), ϕ and θ, are distinct, the inferences based on the likelihood, $L_{obs}(\theta, \phi \mid V_{obs}, R)$,

are equivalent to the inferences from $L_{obs}(\theta \mid V_{obs})$. Hence the missing-data mechanism is "ignorable". In this case one only needs to maximize the $L_{obs}(\theta)$ component of the loglikelihood of the observed data, $L_{obs}(\theta, \phi)$. The second term in the equation (4), $L_{obs}(\phi)$, can be ignored.

However the missing-data mechanism would be "nonignorable" if the process depends on the missing values (V_{miss}) and not on the observed values (V_{obs}) , that is,

$$P(R \mid V_{obs}, V_{miss}, \phi) \neq P(R \mid V_{obs}, \phi). \qquad (5)$$

In this case, the missing data are not missing at random and ML estimation requires a model for the missing data. This occurs because $L_{obs}(\theta, \phi)$ can not be represented as a function of θ plus a function of ϕ : $L_{obs}(\theta, \phi)$ must be maximized jointly with respect to θ & ϕ .

On the other hand, if the missing-data mechanism does not depend on either (V_{obs}) or (V_{miss}) , that is,

$$P(R \mid V_{obs}, V_{miss}, \phi) = P(R \mid \phi), \qquad (6)$$

the missing values are considered missing completely random (MCAR) and the likelihood inferences for the parameter underlying the observed data, θ , can safely disregard the missing-data mechanism. The MCAR cases represent stringent ignorable missing-data mechanism, while less restrictive MAR is usually sufficient for ignorable models.

Many of the analyses that deal with missing-data patterns are conceived under the assumption that the

missing-data mechanism is ignorable. For both univariate and multivariate incomplete data, the missing-data mechanism is ignorable if the missingness is not related to the variables under study (MCAR), or at least not to the missing data (MAR). Rubin notes that ad hoc procedures, such as the complete data method and imputation, are valid under the stringent MCAR assumption. However, for the methods based on the method of Maximum Likelihood, the ignorable missing-data mechanism needs not be MCAR, but weaker MAR assumption will suffice. Both assumptions do not require a model for the missing-data mechanism, which holds practical value.

Under the MAR assumption, the complexity of the ML estimation procedures depend on whether the missing-data pattern is monotonic or nonmonotonic. The full likelihood function for the multivariate incomplete data tends to be complex with many parameters which may or may not be distinct. If the missing-data pattern is monotonic, the maximization of the likelihood is easier since the likelihood factorizes into smaller components whose parameters are distinct for each factor. Each of the factors are maximized separately. Some parameters can be estimated using the standard complete data statistical analyses.

In an example for the Type I missing-data pattern, where there are missing values on only one variable (Figure 1.), the MAR assumption means that, given the observed data in Y_1, whether the values in Y_2 are missing or not do not depend on Y_1. For this type of incomplete data, the MLE's

are derived by specifying the bivariate distribution for V_1 and V_2. The loglikelihood of the joint distribution can be broken down as the product of the marginal density of V_1 and the conditional density for V_2 given V_1;

$$\begin{aligned} L_{obs}(\theta \mid V_{obs}) &= \log P(V_1, V_2 \mid \theta) \\ &= \sum_{i=1}^{n_1} \log P(V_1, V_2 \mid \theta) + \sum_{n_1+1}^{N} \log P(V_1 \mid \theta_{V_1}) \\ &= \sum_{i=1}^{N} \log P(V_1 \mid \theta_{V_1}) + \sum_{i=1}^{n_1} \log P(V_2 \mid V_1, \theta_{V_2 \mid V_1}) . \quad (7) \end{aligned}$$

The maximization of the second term, $L_{obs}(\theta_{V_2 \mid V_1})$ can be separately achieved by performing a regression analysis where V_2 is predicted from V_1 using the n_1 cases who have V_2 scores.

The pattern of multivariate missing-data shown in Figure 2 is also monotonic and is an extension of the Type I missing-data pattern discussed above. If missing data are missing at random and the underlying distribution is multivariate normal, the ML estimation for the Type II missing-data pattern would become a series of regression analyses. In this example with three variables, ML estimation of the mean and covariance matrix involves; first, calculate the mean and covariance for the fully observed V_1 variable; second, perform the linear regression of the next most observed variable V_2 on V_1 for the cases both variables are recorded. Then, calculate the regression of the V_3 on both V_1 and V_2 for the cases in which all three variables are recorded.

When the MAR assumption holds, but the missing-data pattern is not monotonic, the maximization of the likelihood

is complicated. Unlike the monotonic pattern, the loglikelihood cannot be broken into a series of terms, each of which contains different parameters. For example, the loglikelihood for Type III nonmonotonic missing-data pattern is given as;

$$L_{obs}(\theta \mid V_{obs}) = \sum_{i=1}^{n_1} \log P(V_1, V_2 \mid \theta_{V_1 V_2}) + \sum_{n_1+1}^{n_2} \log P(V_2, V_3 \mid \theta_{V_2 V_3}) + \sum_{n_2+1}^{N} \log P(V_1, V_3 \mid \theta_{V_1 V_3}) . \qquad (8)$$

The three terms in the equation (8) represents each of the groups with the parameters, the mean vector and the covariance matrix. However, the parameters, such as means $(\mu_{V_1}, \mu_{V_2}, \mu_{V_3})$ and variances $(\sigma^2_{V_1}, \sigma^2_{V_2}, \sigma^2_{V_3})$, appear more than once in the whole likelihood. One must then maximize the likelihood jointly with respect to all parameters.

The problem described above makes it difficult to maximize the likelihood, $L_{obs}(\theta)$, and obtain estimates when the missing-data pattern is nonmonotonic. In these cases a useful algorithm has been developed for obtaining Maximum Likelihood estimates. Dempster et al. (1977) formalized this procedure as the Expectation-Maximization algorithm. The EM algorithm is a maximization technique which is very useful in finding the estimates for a nonmonotonic missing data pattern. Rubin also notes the broad applicability of the EM algorithm for the models under the nonignorable missing-data mechanism as well as for the ignorable models.

Proposed Method: EM Algorithm

As previously noted, in many situations, the missing-data pattern do not take a monotonic form, in which the factorization of the likelihood leads to the distinct parameters estimable by standard computing methods. The form of the likelihood for the incomplete data with nonmonotonic pattern is "nonstandard", i.e. does not have the form of a complete data likelihood and can not be factored. The EM algorithm is used to maximize the $L_{obs}(\theta \mid V_{obs})$ by estimating the expected values of the hypothetical complete data loglikelihood, $L_{comp}(\theta \mid V_{obs}, V_{miss})$, at each iteration of the algorithm.

The algorithm consists of two steps. First, in E-step, one constructs a quasi complete data loglikelihood. This is accomplished by taking the expected value of the hypothetical complete likelihood, given the observed data. In taking the expectation, one uses some initial guesses for the parameter values. Then, in M-step, one finds those estimates which maximize the expected likelihood constructed in the first step. Using the estimates in the second step, one begins again with the first step. Given that the likelihood is bounded, the algorithm will converge to a solution which is a maximum of the loglikelihood.

The generality of the EM application for obtaining the ML estimates in incomplete data has repeatedly been

demonstrated in research contexts (e.g. McKendrick, 1926; Hartley, 1958; Baum et al, 1970; Orchard & Woodbury, 1972; Sundberg, 1975; Beale & Little, 1975; Dempster, Laird, & Rubin, 1977). Many applications of the EM algorithm have been made in the distributions of incomplete multivariate normal model for continuous variables (e.g. Beale & Little, 1975; Dempster, Laird, and Rubin, 1977; Hartley, 1958; Hartley & Hocking, 1971; Sundberg, 1974; Orchard & Woodbury, 1974), and the multinomial models for the categorical variables in the partially classified contingency tables (e.g. Chen & Fienberg, 1974; Haberman, 1974, Fuchs, 1982). Little & Schulucter (1985) described an integrated model for the multivariate data with incomplete continuous and categorical variables.

The CUNY open-admissions data set is a complicated multivariate incomplete data set with key variables measured in both nominal and interval scales. The present study intends to analyze the data set following the general formulations of Rubin's model based on the missing-data mechanism. In obtaining the Maximum Likelihood estimates for the parameter underlying the distribution of the chosen variables, the EM algorithm suggested by Little & Schulucter was applied.

Chapter IV

STATISTICAL MODEL

As noted in previous chapters, the follow-up study on the original sample for the CUNY open-admissions research has resulted in an enormous amount of data accompanied an extensive missing data problem. The aim of the follow-up study was to assess the means and variances of work and life experience variables, and to investigate the relationship of these follow-up study variables to variables obtained from the original data. The high proportion of missing data came mainly from the fact that only 14 percent of the original sample responded to the follow-up study. Thus there exist no data for 86 percent of the original sample on the variables unique to the follow-up study. Yet, as the respondents to the follow-up study are identified among the original sample, significant amounts of missing values are also found in the original variables as well.

In terms of missing data, the CUNY open-admissions data set is most complicated since it takes on a nonmonotonic missing-data pattern. Generally speaking, if the missing-data pattern is not monotonic, Maximum Likelihood estimation will often require an iterative

solution of the likelihood equations. A further complication is that the data are multivariate and contain both incomplete continuous and categorical variables. Thus a model is needed for the joint distribution of both continuous and categorical variables.

The discussion of a model for mixtures of continuous and categorical variables is a recent development in the literature. Olkin & Tate (1961) defined a multivariate model, in which the categorical variables are multinomial, and the conditional distribution of the continuous variables given the categorical variables is multivariate normal with common variance-covariance matrix. This model is termed the general location model. On the basis of the location model, Krzanowski (1975) proposes a likelihood method for the treatment of mixed binary and continuous variables in discriminant analysis. He noted that the location model is particularly suitable for generalization to mixtures of all types of variables.

Little & Schulucter (1985) based the Maximum Likelihood estimation procedure for the incomplete data with mixed continuous and categorical variables on the general location model of Olkin & Tate and extensions proposed by Krazanowski. The model discussed by Little & Schulucter assumes the missing data are missing at random, complying with Rubin's definition. Hence the likelihood for the model does not require the specification for missing-data mechanism and is regarded as a function of parameters from the joint distribution of continuous and categorical

variables. Little & Rubin (1987) also state that the Maximum Likelihood estimates are particularly useful for various models, such as logistic regression models with missing values, or linear regression models with missing continuous and categorical predictors.

Consider that we have a hypothetical complete data set consisting of a random sample of size n. The following development of notation for the hypothetical data are taken from Little & Rubin (1987, pp.196-197). Suppose the data comprise K continuous variables and V categorical variables. Let I_j denote the number of levels for categorical variable, j. Then the V categorical variables make up V-way contingency table with $C=\prod_{j=1}^{V} I_j$ cells. For each subject i, there is a $1\times K$ vector of continuous variables $(\mathbf{x}_i)$ and a $1\times V$ vector of categorical variables $(\mathbf{y}_i)$. To more easily denote which cell of the contingency table a subject is in, let $\mathbf{u}_i$ be a $1\times C$ binary vector constructed from $\mathbf{y}_i$ in the following manner: If the subject is in cell m, then $\mathbf{u}_i=E_m$ where the vector E_m takes the value 1 in the m_{th} entry and 0's elsewhere.

The symbols and their meanings used to describe the above data are summarized as follows:

n : Sample size

K : Number of continuous variable

V : Number of categorical variable

i : A subject

j : A categorical variable

I_j : Number of levels of categorical variable, j

$C = \prod_{j=1}^{V} I_j$: Cells in the contingency table formed by categorical variables

$\underline{x}_i = (1 \times K)$: Vector of continuous variables

$\underline{y}_i = (1 \times V)$: Vector of categorical variables

$\underline{w}_i = (1 \times C)$: Vector denoting the cell membership of subject i

$E_m = (1 \times C)$: The Same as $\underline{w}_i$ if subject i belongs to cell m of the table comprised of a single 1's & multiple 0's.

The following model is defined by Olkin & Tate for the general location model. For the distributions of $(\underline{x}_i, \underline{w}_i)$, the $\underline{w}_i$ are independent and individually distributed multinomial random variables with cell probabilities;

$$P(\underline{w}_i = E_m) = \pi_m, \quad m = 1, \ldots, C; \Sigma \pi_m = 1. \qquad (1)$$

Given that $\underline{w}_i = E_m$, the conditional distribution of $\underline{x}_i$ given $\underline{w}_i$ is the *K-variate* normal distribution with mean $\mu_m = (\mu_{m1}, \ldots, \mu_{mk})$ and covariance matrix Ω. The cell probabilities are represented by $\Pi = (\pi_1, \ldots, \pi_c)$ for the $(1 \times C)$ vector. The cell means are represented by $\Gamma = \{\mu_{mk}\}$ for the $(C \times K)$ matrix. The parameters in the

model are represented by $\Theta = (\Pi, \Gamma, \Omega)$. There are $C - 1 + KC + \frac{1}{2}K(K+1)$ parameters to be estimated.

The complete-data loglikelihood for the location model is stated Little & Rubin (1987) as follows:

$$\begin{aligned} l(\Gamma, \Omega, \Pi) &= \sum_{i=1}^{n} \ln f(x_i \mid w_i, \Gamma, \Omega) + \sum_{i=1}^{n} \ln f(w_i \mid \Pi) \\ &= h(\Omega) - \frac{1}{2} tr\left((\Omega^{-1} \sum_{i=1}^{n} x_i^T x_i \right) + tr\,\Omega^{-1}\Gamma\left(\sum_{i=1}^{n} w_i^T x_i \right) \\ &\quad + \sum_{m=1}^{C} \left[\left(\sum_{i=1}^{n} W_{im} \right) (\ln \pi_m - \frac{1}{2} \mu_m \Omega^{-1} \mu_m^T) \right], \end{aligned} \tag{2}$$

where w_{im} is the m_{th} component of w_i , tr means "trace of the matrix," and $h(\Omega) = -\frac{1}{2} n \{K \ln(2\pi) + \ln(|\Omega|)\}$. Maximizing (2) yields complete-data ML estimates;

$$\begin{aligned} \hat{\Pi} &= n^{-1} \sum w_i, \\ \hat{\Gamma} &= (\sum x_i^T w_i)(\sum w_i^T w_i)^{-1}, \\ \hat{\Omega} &= n^{-1} \sum (x_i - w_i \hat{\Gamma})^T (x_i - w_i \hat{\Gamma}), \end{aligned} \tag{3}$$

which are simply the observed cell proportions, the observed cell means, and the pooled within-cell covariance matrix of X, respectively.

Simple Example of EM Algorithm for 1 Continuous and 1 Categorical Variable

As a simple example, the Maximum Likelihood procedure for the model above can be illustrated with n observations

where each subject is measured on one continuous variable and one categorical variable. Suppose the categorical variable has four levels, $a, b, c, \& d$. Each level is represented by cell probabilities $\Pi = (\pi_a, \pi_b, \pi_c, \pi_d)$. In each level of the categorical variable, a continuous variable is normally distributed with mean μ_m ; $m = a, b, c, \& d$, and variance σ^2. Thus, there are eight parameters to be estimated, $\pi_a, \pi_b, \pi_c, \mu_a, \mu_b, \mu_c \mu_d \& \sigma^2$.

According to the general location model, the complete data model consists of the joint distribution of $(x, \mathbf{w})$, for n observations. Since there are four levels for the categorical variable, the possible values for $\mathbf{w}_i$ are; $\mathbf{w} = [1000], [0100], [0010], [0001]$. The marginal distribution of $\mathbf{w}$ is a multinomial distribution over the three cells of the contingency table. The conditional distribution of x given $\mathbf{w} = E_m$ is a normal distribution with mean μ_m and variance σ^2.

Suppose the data have the following missing-data pattern (R):

pattern	X	$\mathbf{W}$	R_x	R_w
1	x	$\mathbf{w}$	1	1
2	x	?	1	0
3	?	$\mathbf{w}$	0	1 ,

where pattern 1 shows complete data for both continuous and categorical variables. For the subjects in pattern 2, the information regarding cell membership is not available. In the third pattern, though we know the cell membership, the

values for the continuous variables are missing. The loglikelihood of the observed data, assuming missing-data are missing at random, will be;

$$L_{obs} = \sum_{grp1} \log p(\mathbf{M})\, p(\mathbf{x} \mid \mathbf{M}) + \sum_{grp2} \log p(\mathbf{x}) + \sum_{grp3} \log p(\mathbf{M}). \qquad (4)$$

For this presentation, instead of directly maximizing the observed-data loglikelihood, we consider the EM algorithm for ML estimation based on the theoretical complete-data loglikelihood:

$$L_{comp} = \sum_{i=1}^{n} \log \{ p(\mathbf{M}_i)\, p(\mathbf{x}_i \mid \mathbf{M}_i) \} = \sum_{i=1}^{n} \log p(\mathbf{M}_i) + \sum_{i=1}^{n} \log p(\mathbf{x}_i \mid \mathbf{M}_i), \qquad (5)$$

If we rewrite (5) using the general form given in (2), it can be seen that the complete-data loglikelihood is linear in a set of sufficient statistics. These statistics are: n_a , the number of people in level a ; n_b , the number of people in level b ; n_c , the number of people in level c ; Σx_a , the sum of the x's in level a ; Σx_b , the sum of the x's in level b ; Σx_c , the sum of the x's in level c ; Σx_d , the sum of the x's in level d ; and Σx^2 , the total sum of the squares for the x's.

The EM algorithm for this model consists of the following steps:

(1) Define the initial estimates of the parameters; $\pi_a, \pi_b, \pi_c, \mu_a, \mu_b, \mu_c, \mu_d, \sigma^2$.

(2) Compute the expected value of the sufficient statistics $n_a, n_b, n_c, \Sigma x_a, \Sigma x_b, \Sigma x_c, \Sigma x_d, \Sigma x^2$, given both observed data (obs) and initial estimates (I) of the parameters.

(3) Replace complete-data sufficient statistics by expected values in (2).

(4) Using (3), get Maximum Likelihood estimates of the parameters. These estimates are as follows: the cell probabilities, $\hat{\pi}_a = \hat{n}_a / n$, $\hat{\pi}_b = \hat{n}_b / n$, and $\hat{\pi}_c = \hat{n}_c / n$; the means, $\hat{\mu}_a = \Sigma \hat{x}_a / n_a$, $\hat{\mu}_a = \Sigma \hat{x}_a / n_a$, $\hat{\mu}_c = \Sigma \hat{x}_c / n_c$, and $\hat{\mu}_d = \Sigma \hat{x}_d / n_d$; the variance, $\hat{\sigma}^2 = \Sigma \hat{x}^2 - n_a \hat{\mu}_a^2 - n_b \hat{\mu}_b^2 - n_c \hat{\mu}_c^2 - n_d \hat{\mu}_d^2 \; / \; n$. These estimates replace the values in (1). The " $\hat{}$ " notation refers to expected values.

Consider first, the computation of the expected values for $n_a, n_b, n_c, \& n_d$. Suppose the m_{th} entry of $\mathbf{w}_i$ is denoted as w_{im} . Then $\mathbf{w}_i = [w_{ia} w_{ib} w_{ic} w_{id}]$. In our example with one categorical and one continuous variable, the sufficient statistics $n_a, n_b, n_c, \& n_d$ are represented as $\Sigma \mathbf{w}_i = [n_a n_b n_c n_d]$. To illustrate the estimation procedure, consider the expected value of n_a , the number of people in level a . Patterns 1 and 3 contain a known number of people who belong to level a , but the membership is not known in pattern 2. The expected value of n_a , given observed data (obs) and initial estimates (I) of parameters, is given as:

$$E(n_a \mid obs, I) = n_{1a} + E(n_{2a} \mid obs, I) + n_{3a}. \qquad (6)$$

n_{1a} and n_{3a} are the observed numbers of people in patterns 1 and 3 who are in cell a, and

$$E(n_{2a} \mid obs, I) = \sum_2 E(w_{ia} \mid obs, I)$$
$$= \sum_2 p(w_{ia} = 1 \mid x_i, I). \qquad (7)$$

In another words, for each subject in pattern 2, we compute the probability of being in cell a, given the x score and the initial estimates. This probability is given as:

$$p(a)\,p(x_i \mid a) / p(x_i)$$
$$= \hat{\pi}_a f(x_{ia}; \hat{\mu}_{ia}, \hat{\sigma}^2) / \sum_{m=a}^{d} \hat{\pi}_m f(x_{im}; \hat{\mu}_{im}, \hat{\sigma}^2), \qquad (8)$$

where $f(x_{im}; \hat{\mu}_{im}, \hat{\sigma}^2)$ is the normal density for x_i with mean, $\hat{\mu}_m$ (initial estimate); variance, $\hat{\sigma}^2$ (initial estimate); and $\hat{\pi}_m$ is the initial estimate of the probability of being in cell m. The method described above is applied to obtain estimates for the number of people in cells, $b, c, \& d$

Next, consider obtaining the expected values of the sufficient statistics, $\Sigma x_a, \Sigma x_b, \Sigma x_c, \Sigma x_d$. Again, let w_i be the vector indicating the cell membership of subject i. With a single categorical variable, the four cell sums of the x's can be described as $\Sigma x_i w_i = [\Sigma x_a \Sigma x_b \Sigma x_c \Sigma x_d]$. The expected value of the sum of the x's in the level a is given as;

$$E(\Sigma x_{ia} \mid obs, I) = \sum_1 x_{ia} + \sum_2 E(x_{ia} \mid x_i, I) + \sum_3 E(x_{ia} \mid w_i, I). \qquad (9)$$

For the second term in (9), the expected values are computed

as;

$$\sum_{2} E(x_{ia} \mid x_i, I) = \sum_{2} E(x_i w_{ia} \mid x_i, I)$$
$$= \sum_{2} x_i p(w_{ia} = 1 \mid x_i, I), \qquad (10)$$

where all x's are observed and $p(w_{ia}=1 \mid x_i, I)$ is obtained using (8). For the last term in (9), the summation is over all cases in pattern 3 where $w_i = [1000]$. For these cases, the expected values are given as $E(x_{ia} \mid w_i, I) = \beta_a$. The same method is used to obtain estimates for the x sums in cells, $b, c, \& d$.

The expected value of the sufficient statistic Σx^2 is given as;

$$E(\Sigma x^2 \mid obs, I) = \sum_{1} x^2 + \sum_{2} x^2 + \sum_{3} E(x^2 \mid w_i, I), \qquad (11)$$

where each missing x^2 in pattern 3 is replaced with its expected value, $\beta_m^2 + \theta^2$, where $m = a, b, c, \& d$ depending on the values of w.

Chapter V

METHOD

The major question of interest in the CUNY open admissions data set is whether open admissions has benefitted the recipients socioeconomically. To investigate this question, Lavin's follow-up study considered questions regarding education attainments as of 1984; labor market experiences including salary for the job held in 1984; and satisfaction with current life situations pertaining to family life, community life, and life in general. These variables are only observed in the follow-up study and are missing for the subjects in the original sample who are not in the follow-up sample.

The following two variables were chosen in the present study as an example of categorical variable and continuous variable, respectively:

SATLIFE — Satisfaction with Life in general: categorical variable with 2 levels; Very Satisfied is categorized as 'Yes' satisfied; somewhat satisfied or not satisfied as 'None to Somewhat' satisfied.

SALNOW — Current Salary: continuous variable, salary before taxes of current job

Also, the variables of major interest in the original sample used for the present study are:

DEGREEA	Degrees aspired: categorical variable with 3 levels; a high school diploma and associate degree (AA) are indexed together in 'None-AA' category; bachelor's degree (BA) in 'BA' category; master's (MA) and beyond masters, in 'MA-Beyond' category.
SESSEXA	Gender: a categorical variable.
RACEA	Race/Ethnic: four ethnic categories, Black, Hispanic, Asian/other, and White, are collapsed into two categories; 'Majority' which is White and 'Minority' that includes all other ethnicity.
CAA	High School Average: a continuous measure of the student's grades in college preparatory courses (e.g., English, mathematics, science).
LASTGPA	Grade Point Average: a continuous variable.
TOTUNITS	Total Units of academic courses: total number of college preparatory courses taken, e.g., English, mathematics, sciences, foreign language, history, and social studies, measured on a continuous scale.

These variables were chosen, since in Lavin's previous analyses, there was some evidence that the respondents and nonrespondents differed on these variables.

Table 1 shows the percentage of missing cases for both original and follow-up variables. Due to the limitation in computer time, the size of the original sample for the present study is N = 4,992. This represents the 14.5 % random sample taken from the total original sample of N = 34,434. This percentage was chosen since it yielded an approximate original sample of 5,000. In this present sample of N = 4,992, N = 4,282 did not respond to the follow-up survey. The follow-up sample of N = 710 was identified from this random sample of N = 4,992 by the variable that indicates their response to the follow-up questionnaire.

Table 1

Percentages of Missing Cases for Original and Follow-up Sample Variables

Variables	Original Sample (N=4,992)	Follow-up Sample (N=710)
ORIGINAL SAMPLE VARIABLES		
DEGREEA	26.2 %	21.1 %
(Valid Cases)	3,684	560
(Missing Cases)	1,308	150
SESSEXA	0 %	0 %
(Valid Cases)	4,992	710
(Missing Cases)	0	0
RACEA	2.8 %	2.1 %
(Valid Cases)	4,852	695
(Missing Cases)	140	15
CAA	5.0 %	2.5 %
(Valid Cases)	4,743	692
(Missing Cases)	249	18
LASTGPA	6.6 %	4.1 %
(Valid Cases)	4,663	681
(Missing Cases)	329	29
TOTUNITS	5.5 %	2.5 %
(Valid Cases)	4,718	692
(Missing Cases)	274	18
FOLLOW-UP SAMPLE VARIABLES		
SATLIFE	86.0 %	1.3 %
(Valid Cases)	701	701
(Missing Cases)	4,291	9
SALNOW	87.7 %	13.4 %
(Valid Cases)	615	615
(Missing Cases)	4,377	95

The percentage of missing cases are computed from the sample size actually used in present analysis. The original variables which are common to both original sample and follow-up sample show a similar proportion of missingcases in both samples. For example, DEGREEA, which measures the level of students' aspiration toward degree, is missing 26.2% in the original sample. Follow-up sample shows similar proportion of missing cases, 21.1%, for the same variable.

The proportion of missing cases for the follow-up variable is the entire subjects (N = 4,282) who did not respond to the follow-up questionnaire as well as a few subjects who did not answer this particular variable among the follow-up sample. For example, SATLIFE, which measures the respondent's satisfaction with life in general, is missing entire 4,282 cases who did not respond to the follow-up survey. In addition, it also contains 9 people who did not answer that particular question among follow-up survey respondents. Therefore the missing-data are present in both original variables as well as the follow-up variables which are measured both on a continuous and categorical scale.

For the analysis of the original variables, that is, in obtaining the means and the variance-covariance matrix of the variables of interest in the original sample with missing data, the following types of missing data treatment are possible: pairwise deletion, listwise deletion, and Maximum Likelihood estimation. Both the pairwise and the

listwise deletion are ad-hoc treatments of missing data and are widely used through statistical packages. The Maximum Likelihood estimation procedure is theory-based. It was implemented in the present study using EM algorithm suggested by Little and Schulucter (1985).

The pairwise deletion excludes missing values pairwise for cases missing one or both members of a pair of variables from the analysis. For analyses which focus on a set of variables "one at a time", e.g. computing means, the pairwise deletion strategy uses a different sample size for each variable. The listwise deletion excludes cases missing on any variable named in the variable list from all analyses. The Maximum Likelihood estimation provides a general method for constructing statistical estimates for population parameters from sample data.

To maximize the likelihood of the observed data with missing values, the EM algorithm, first, obtains the initial estimates of the parameters of missing data based on the observed complete data. Then, the algorithm computes the expected value of the sufficient statistics given both observed data and the initial estimates of the parameters. Thus, the unobserved compete-data sufficient statistics are replaced by their expected values given the observed data and initial estimates of the parameters. Using these readily expected values, maximum likelihood estimates are obtained and the cycle begins all over again until the estimates converge. A simple example of EM algorithm for 1 continuous and 1 categorical was considered in Chapter IV.

For the current analysis comparing the parameters estimated from the three different types of analysis, the number and types of variables which could be studied was limited by the computational requirements of the Maximum Likelihood method. For this analysis, a FORTRAN program provided by Schulucter (1985) was utilized. The program is set up to run up to 8 continuous variables, 4 categorical variables, with up to 32 cells in the contingency table, and a maximum of 7 levels per categorical variable. Further, the sample size is limited by memory requirements. The eight variables listed above: satisfaction with life (SATLIFE), salary as of 1984 (SALNOW), high school average (CAA), degree aspired (DEGREEA), gender (SESSEXA), race (RACEA), mean GPA (LASTGPA), mean total academic units (TOTUNITS), were selected for all of three types of analysis. There are four categorical variables; DEGREEA, SESSEXA, RACEA, SATLIFE, which constitutes 3 X 2 X 2 X 2 contingency table, and four continuous variables; CAA, LASTGPA, SALNOW, and TOTUNITS.

In order to replicate previous analyses of this data set, a fourth method of analysis was used. In this ad hoc procedure, an attempt was made to make the follow-up sample representative of the original sample. This involves predicting the likelihood that a given individual in the original sample would have responded to the follow-up questionnaire, based on what we know about the individual's original sample characteristics. Lavin's follow-up study used this weighting procedure to make inferences about an

original sample (N = 34,234) based on the conclusions drawn from the follow-up sample (N = 4,989). Their weighting procedure was based on a strategy suggested by Berk (1983). Lavin et al. found a number of variables in the original sample that were useful in predicting the odds of being in the follow-up sample. These included race, gender, age, income, high school average, entry cohort, level of entry to CUNY (senior or community college), number of credits earned at CUNY, and graduation from CUNY.

In the present study, only three of these original sample variables were used, due to the restriction in the number and level of categorical variables allowed for by the computer program used to obtain the Maximum Likelihood estimates. The two categorical variables chosen were race (RACEA) and gender (SESSEXA), and the one continuous variable was high school average (CAA). In addition, two continuous (LASTGPA and TOTUNITS) and one categorical (DEGREEA) original sample variables were chosen for the weighting and other types of analysis. They were selected because they are related to educational and socioeconomic outcome.

Therefore, for the weighting analysis in the present study, where the dependent variable is the log odds that someone from the original sample (N = 4,992) would respond to the follow-up variables, a logistic regression equation was created using six original variables; RACEA, SESSEXA, CAA, LASTGPA, TOTUNITS, and DEGREEA. After estimating the logistic regression equation, the log odds were converted

into a probability. The inverse of this probability is used to weight individuals (N = 710) in the follow-up sample.

Chapter VI

RESULTS

In Tables 1-3, estimates from the various types of analyses are presented. Table 1 shows estimates of proportions for the three original sample categorical variables (DEGREEA, RACEA, & SESSEXA). Estimates of the means and standard deviations of the three original sample continuous variables (CAA, LASTGPA, & TOTUNITS) are presented in Table 2. Table 3 contains the estimated proportions for the follow-up categorical variable (SATLIFE), and the mean and standard deviations for the follow-up continuous variable (SALNOW). The correlation coefficients between the original sample and the follow-up sample continuous variables are illustrated in Table 4. In each of the first three tables, three different methods were used to analyze the data: Pairwise deletion, Listwise deletion and Maximum Likelihood estimation. For the follow-up variables, an additional weighting analysis was added.

As is evident throughout the tables, there are large differences in the number of valid cases utilized in each type of analysis. For example, in columns 2 and 3 in

Table 1, the number of valid cases, N = 3,864 for the pairwise deletion analysis of the DEGREEA variable is reduced considerably to N = 455 when the listwise deletion method of missing data was employed. These differences in the number of cases used are not as great for the follow-up sample variables. In Table 3, for the follow-up sample variable SATLIFE, the valid cases used in the pairwise deletion is N = 701 while the valid cases for the listwise deletion is N = 455. In the follow-up variables, although there are some missing data present among follow-up study participants, most of the missing cases are the subjects who did not respond to the follow-up questionnaire at all. The valid cases for the weighting analysis represents the sum of all weighted cases. Further, in all tables, it should be noted that since the Maximum Likelihood method using EM algorithm utilizes all available data, it is not possible to give the valid number of cases used in the estimates of each variable.

For the original sample variables, since there are few or no missing data, little differences were expected among the estimates using three methods of analyzing missing data. In both Tables 1 and 2, any differences between estimates is a function of the proportion of missing cases. For example, for the variable LASTGPA with missing cases of 140, Table 1 shows that the estimates from these two types of analysis, Pairwise deletion and Maximum Likelihood, are 72.0 and 71.8 respectively. On the other hand, there is a considerable difference between these estimates in the variable DEGREEA

Table 1

Estimates of Proportions for Original Sample Categorical Variables: A Comparison of Pairwise Deletion, Listwise Deletion, and Maximum Likelihood Estimates

Variables	Pairwise Deletion	Listwise Deletion	Maximum Likelihood Estimates
ORIGINAL SAMPLE CATEGORICAL VARIABLES			
DEGREEA(%)			
none-AA	16.2	11.9	17.9
BA	34.4	34.3	34.3
MASTERS-beyond	49.5	53.8	47.8
(Valid Cases)	(3,684)	(455)	*
(Missing Cases)	(1,308)	(4,537)	*
RACEA(%)			
majority	72.0	74.9	71.8
minority	28.0	25.1	28.2
(Valid Cases)	(4,852)	(455)	*
(Missing Cases)	(140)	(4,537)	*
SESSEXA(%)			
male	48.6	47.3	48.6
female	51.4	52.7	51.4
(Valid Cases)	(4,992)	(455)	*
(Missing Cases)	(0)	(4,537)	*

* ML methods use all available data.

where there are more than 1,000 cases missing. The estimate for DEGREEA using Pairwise deletion of missing data in the 'None-AA' category is 16.2 while the Maximum Likelihood Estimate for the same category is 17.9. The estimates using the Listwise deletion method differ most from both the Pairwise deletion and the Maximum Likelihood methods due to the greater number of missing data.

Table 2

Estimates of Means and Standard Deviations for Original Sample Continuous Variables: A Comparison of Pairwise Deletion, Listwise Deletion, and Maximum Likelihood Estimates

Variables	Pairwise Deletion	Listwise Deletion	Maximum Likelihood Estimates
ORIGINAL SAMPLE CONTINUOUS VARIABLES			
CAA			
Mean	77.35	79.98	77.26
SD	7.78	7.29	6.88
(Valid Cases)	(4,743)	(455)	*
(Missing Cases)	(249)	(4,537)	*
LASTGPA			
Mean	2.27	2.56	2.25
SD	.89	.76	.83
(Valid Cases)	(4,663)	(455)	*
(Missing Cases)	(329)	(4,537)	*
TOTUNITS			
Mean	12.39	13.43	12.34
SD	3.12	2.50	2.30
(Valid Cases)	(4,718)	(455)	*
(Missing Cases)	(274)	(4,537)	*

* ML methods use all available data.

The larger discrepancies in estimates among different types of missing data analyses are present in the analysis of Follow-up sample variables. Table 3 shows that for the variable SATLIFE, the estimate using the Pairwise deletion is 55.6 for the 'Yes' category. On the other hand, the Maximum Likelihood Estimate is 54.9 in the same category. For the variable SALNOW, the mean estimate using the pairwise deletion is 25,010.79 whereas the Maximum Likelihood estimate is 24,096.89. The estimates using

Table 3

Estimates of Proportions, Mean, and Standard Deviation for Follow-up Sample Variables: A Comparison of Pairwise Deletion, Listwise Deletion, Weighting Analysis, and Maximum Likelihood Estimates

Variables	Pairwise Deletion	Listwise Deletion	Weighting Analysis	Maximum Likelihood Estimates
FOLLOW-UP SAMPLE CATEGORICAL VARIABLE				
SATLIFE(%)				
yes	55.6	57.8	55.2	54.9
none-some	44.4	42.2	44.8	45.1
(Valid Cases)	(701)	(455)	(4,992)**	*
(Missing Cases)	(4,291)	(4,537)	(0)**	*
FOLLOW-UP SAMPLE CONTINUOUS VARIABLE				
SALNOW(%)				
Mean	25,010.79	25,903.58	24,097.53	24,096.89
SD	15,395.76	18,048.54	15,099.39	15,189.91
(Valid Cases)	(615)	(455)	(4,992)**	*
(Missing Cases)	(4,337)	(4,537)	(0)**	*

* ML methods use all available data.
** Valid cases of 4,992 is the sum of all weighted cases, i.e. sum of the reciprocals of the probability of being in the follow-up sample for the actural respondents.

Listwise deletion method of missing data continue to differ from both Pairwise deletion and Maximum Likelihood estimation in the analysis of the follow-up sample variables. For the same 'Yes' category in the variable SATLIFE, the Listwise deletion estimate is 57.8. The estimates for the SALNOW variable is 25,903.58 using the Listwise deletion.

The estimates of the mean, the standard deviation, and proportions for the follow-up sample variables using the

Table 4

Correlation Coefficients for Continuous Variables: A Comparison of Listwise Deletion, Pairwise Deletion and Weighting Analysis

LISTWISE DELETION

	CAA	LASTGPA	SALNOW	TOTUNITS
CAA	1.00 (N=577)			
LASTGPA	0.52 (N=577)	1.00 (N=577)		
SALNOW	0.07 (N=577)	0.13 (N=577)	1.00 (N=577)	
TOTUNITS	0.57 (N=577)	0.40 (N=577)	0.15 (N=577)	1.00 (N=577)

PAIRWISE DELETION

	CAA	LASTGPA	SALNOW	TOTUNITS
CAA	1.00 (N=4,743)			
LASTGPA	0.49 (N=4,470)	1.00 (N=4,663)		
SALNOW	0.08 (N=598)	0.13 (N=594)	1.00 (N=615)	
TOTUNITS	0.61 (N=4,697)	0.36 (N=4,444)	0.16 (N=598)	1.00 (N=4,718)

PAIRWISE DELETION USING WEIGHTED FOLLOW-UP SAMPLE VARIABLE

	CAA	LASTGPA	WSALNOW*	TOTUNITS
CAA	1.00 (N=4,743)			
LASTGPA	0.49 (N=4,470)	1.00 (N=4,663)		
WSALNOW*	0.23 (N=598)	0.33 (N=594)	1.00 (N=615)	
TOTUNITS	0.61 (N=4,697)	0.36 (N=4,444)	0.32 (N=598)	1.00 (N=4,718)

MAXIMUM LIKELIHOOD ESTIMATES**

	CAA	LASTGPA	SALNOW	TOTUNITS
CAA	1.00			
LASTGPA	0.42	1.00		
SALNOW	0.08	0.19	1.00	
TOTUNITS	0.56	0.29	0.12	1.00

* weighted SALNOW follow-up sample variable
** ML methods use all available data.

weighting analysis are presented in Table 3. There are only slight differences between the weighting analysis estimates and the estimates using Maximum Likelihood as well as Pairwise deletion. For example, for the 'Yes' category in the SATLIFE variable, the estimate using the weighting analysis is 55.2 whereas the Pairwise deletion method estimate is 55.6 and the Maximum Likelihood is 54.9. The weighting procedure achieves its estimates by weighting each individual with the probability that any given individual in the original sample would responded to the follow-up sample based on the known characteristics of that individual in the original sample.

The biggest differences in the estimates can be seen in the correlations of the original sample and follow-up continuous variables in Table 4. In this table, the correlation coefficients of TOTUNITS with CAA, LASTGPA, and SALNOW are .57, .40, and .15 respectively using Listwise deletion. The correlations of the same variables using the Pairwise deletion are .61, .36, and .16. The Maximum Likelihood estimates of these correlations are .56, .29, and .12. When the original sample variable, such as TOTUNITS, is correlated with the weighted follow-up sample variable, the estimates are .61, .36, and .32. It should be noted that the coefficient became almost twice as large as .32 when TOTUNITS is correlated with the weighted follow-up sample variable, WSALNOW (i.e. weighted SALNOW).

In addition, the multiple correlation of the follow-up variable, SALNOW, with original variables, CAA, LASTGPA,

TOTUNITS, was completed for the four correlation matrices in Table 4. The multiple correlation for Listwise, Pairwise, and Maximum Likelihood methods were quite similar, being in the range from .15 to .20. However, the value for the Pairwise deletion method using a weighted SALNOW variable was much higher, attaining a value of .40. The reason for this discrepancy seems to due to an inflation of the correlation between the dependent variable and the GPA variable, LASTGPA. The reason for this result is unclear.

Chapter VII

SUMMARY AND DISCUSSION

The present study utilized a sample of the data set from the CUNY open-admissions follow-up study. Due to limitations in computer time, the size of the original sample for the current study consists of 14.5% random sample from Lavin's original sample. The choice of variables used in the current study were also restricted by computer time, resulting in the selection of a total eight variables. Data were not only missing for the follow-up variables but also for the original sample variables.

An original sample consisting of N = 4,992 cases was measured on three categorical variables; DEGREEA (Degree Aspired), SESSEXA (Gender), RACEA (Race/Ethnicity), three continuous variables; CAA (High School Average), LASTGPA (Grade Point Average), TOTUNITS (Total Units of Academic Courses). The follow-up sample comprises N = 710 of the original sample. The follow-up sample variables are a categorical variable, SATLIFE (Satisfaction with Life in General) and a continuous variable, SALNOW (Current Salary). The descriptions of each variable in detail are contained in Chapter V.

The purpose of the present study is to analyze the data set under the more realistic assumption that the missing data are missing at random (MAR) rather than missing completely at random (MCAR). The EM algorithm was used to obtain the Maximum Likelihood estimates of the means, standard deviations and correlations for the continuous variables, and proportions for the categorical variables. The Maximum Likelihood estimates were compared with other widely used methods of missing data analysis, i.e. Pairwise deletion, Listwise deletion, and a Weighting method. The means and standard deviation for continuous variables and the proportions for the categorical variables were shown in Tables 1 through Table 3 of Chapter VI. Correlations between the original sample continuous variables and the follow-up sample continuous variable are also presented in Table 4 in Chapter VI.

Despite some noticeable differences in correlations among continuous variables, the overall results in Chapter VI show little evidence that the different methods of missing data treatment yield different estimates for the eight variables used in the current study. The fact that there is a small difference between the Maximum Likelihood estimates and the Pairwise and Listwise deletion methods suggests that the missing data are missing completely random (MCAR). In other words, the missing data mechanism does not depend on either the missing data or the observed data.

Evidence indicating that respondents do not differ from the nonrespondents, i.e. the follow-up data are missing

completely at random, can be drawn from the results of the logistic regression analysis used in the Weighting method. Weighting procedures involved producing weights which are the inverse of the probability of a subject being in the follow-up sample, given the scores on the original sample variables. As a part of weighting analysis, these weights were obtained from a logistic regression analysis. The original sample variables used in the construction of a logistic multiple regression model are; DEGREEA, SESSEXA, RACEA, CAA, LASTGPA, & TOTUNITS.

It can be argued, that if the follow-up missing data are missing completely random, the logistic regression equation should not differentiate the respondents to the follow-up sample from the non-respondents using the above six original sample variables. Appendix A provides the estimated coefficients and related statistics from the logistic regression model. It also provides a confusion matrix which shows how well the logistic regression model fits the data. In the table, the two rows represent whether or not a subject in the original sample is in the follow-up sample; the two columns describe the predicted results. When the Cohen's Kappa (1988) was computed to evaluate the accuracy of classification, the results show a near zero Kappa which suggests that the model does not differentiate between people who responded to the follow-up survey and those who do not in terms of the original sample variables.

These results support the argument that the missing data are missing completely at random (MCAR). Given MCAR,

one would not expect differences between the estimation methods. However, one could also argue that the data are missing at random (MAR) and not MCAR. An inspection of the logistic regression on page 63, in Appendix A, shows that some regression coefficients were significantly different from zero, and further the predicted probabilities of being in the follow-up sample were not constant for all cases. Therefore the MCAR argument cannot completely explain the results. A more plausible explanation is that the methods produced similar results due to the extremely high percentages of missing data, approximately 86%. Under less extreme conditions, different results would be expected.

Thus the results of the present study should not be taken as an indication that the various methods of missing data treatment can be used interchangeably. In general the Maximum Likelihood method is most desirable since, unlike the other procedures, it is based on a well defined statistical model. However, it should be noted that the EM program was extremely computationally expensive; its execution required enormous computer time. Consequently, the number of variables and, especially, the number of cells in the categorical variables allowed in the program has to be limited to avoid a possible failure to converge.

Future Research

A number of approaches might be used to examine the effect that the amount of missing data has on the different

methods of missing data treatment. One way to check the effect of varying proportion of missing data in the CUNY Open-admissions data may be to compare estimates when all 4,988 cases in Lavin et al's follow-up sample are used and the different amount of original sample who are missing in the follow-up variables are added in the analyses. However, the Maximum Likelihood method is still restricted by the computer time and resources.

Since an iterative procedure using a computer program can be quite expensive, another way to analyze the CUNY Open-admissions study missing data would be to use the noniterative estimation procedure. Like many other follow-up studies, the pattern of missing data in the CUNY Open-admissions data is approximately nested; that is, the subjects in the original sample are measured on almost all original variables and the subjects in the original sample who did not respond to the follow-up study are missing values on all follow-up variables.

The noniterative procedure involves arranging the variables that are most observed as a block 1, then the next most observed variables as block 2, and so on. The Maximum Likelihood estimates of the mean and covariance matrix of the variables can be found as follows: (1) Calculate the mean and the covariance of block 1. (2) Calculate the multiple linear regression of the next most observed variables, block 2 on block 1, from observation with both block 1 and block 2 variables measured (Little & Rubin, 1986).

Appendix A

THE ESTIMATED COEFFICIENTS, RELATED STATISTICS, AND A CLASSIFICATION TABLE FROM LOGISTIC REGRESSION ANALYSIS

---------- CORRELATION COEFFICIENTS ----------

	SPENSAMP	RACEA	CAA	TOTUNITS	LASTGPA	SESSEXA
SPENSAMP	1.0000	.0237	.0756**	.0904**	.0717**	.0205
RACEA	.0237	1.0000	.0558**	.0475**	.0202	-.0093
CAA	.0756**	.0558**	1.0000	.8046**	.2365**	.0519**
TOTUNITS	.0904**	.0475**	.8046**	1.0000	.2527**	-.0133
LASTGPA	.0717**	.0202	.2365**	.2527**	1.0000	.0204
SESSEXA	.0205	-.0093	.0519**	-.0133	.0204	1.0000

* - SIGNIF. LE .05 ** - SIGNIF. LE .01 (2-TAILED)
" . " IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED

---------- CLASSIFICATION TABLE FOR FOLLOW-UP SAMPLE ----------

		PREDICTED NOT IN FOLLOW-UP N	PREDICTED FOLLOW-UP SAMPLE I	PREDICTED S	PERCENT CORRECT
OBSERVED					
NOT IN FOLLOW-UP	N	4282	0		100.00%
FOLLOW-UP SAMPLE	S	710	0		.00%
				OVERALL	85.78%

---------- VARIABLES IN THE EQUATION ----------

VARIABLE	B	S.E.	WALD	DF	SIG	R	EXP(B)
MINORD	.4004	.0981	16.6611	1	.0000	.0599	1.4925
UNIT2D	.4306	.1978	4.7374	1	.0295	.0259	1.5382
UNIT3D	.6292	.1535	16.7972	1	.0000	.0602	1.8762
UNIT4D	.8765	.1536	32.5453	1	.0000	.0865	2.4024
GPA5D	.4650	.1040	20.0030	1	.0000	.0664	1.5920
GPA6D	.5717	.1048	29.7709	1	.0000	.0825	1.7712
CONSTANT	-2.8023	.1487	355.0398	1	.0000		

---------- OBSERVED GROUPS AND PREDICTED PROBABILITIES ----------

```
F        I                                                        I
R   1200 +                                  SYMBOLS: N - NOT IN FOLLOW-UP+
E        I                                           S - FOLLOW-UP SAMPLEI
Q        I                                 EACH SYMBOL REPRESSENT 100 CASESI
U        I       S                                                 I
E    800 +       N                                                 +
N        I      NN    S                                            I
C        I      NN    N                                            I
Y        I      NN   SN                                            I
     400 +      NNS  NN                                            +
         I   NNNNNNSNNN                                            I
         I   NNNNNNNNNN                                            I
         I   NNNNNNNNNN  NS                                        I
PREDICTED ------------+-------------+-------------+---------------
  PROB:   0            .25           .5            .75             1
  GROUP:  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNSSSSSSSSSSSSSSSSSSSSSSSSSSSS
          PREDICTED PROBABILITY IS OF MEMBERSHIP FOR FOLLOW-UP SAMPLE
```

Appendix B

PROPORTIONS AND MEANS FOR ORIGINAL SAMPLE VARIABLES: A COMPARISON BETWEEN FOR CASES NOT IN FOLLOW-UP SAMPLE AND FOR CASES IN FOLLOW-UP SAMPLE

Proportions and Means for Original Sample Variables:
A Comparison between for Cases not in the Follow-up Sample
and for Cases in the Follow-up Sample

Original Sample Variables	For Cases Not In the Follow-up Sample	For Cases In the Follow-up Sample
ORIGINAL SAMPLE CATEGORICAL VARIABLES		
DEGREEA(%)		
none-AA	16.8	12.3
BA	34.6	33.4
MASTERS-beyond	48.6	54.3
(Valid Cases)	(3,124)	(560)
(Missing Cases)	(1,158)	(150)
RACEA(%)		
majority	72.2	71.1
minority	27.8	28.9
(Valid Cases)	(4,157)	(695)
(Missing Cases)	(125)	(15)
SESSEXA(%)		
male	49.0	46.1
female	51.0	53.9
(Valid Cases)	(4,282)	(710)
(Missing Cases)	(0)	(0)
ORIGINAL SAMPLE CONTINUOUS VARIABLES		
CAA(%)		
Mean	77.06	79.04
SD	7.8	7.4
(Valid Cases)	(4,051)	(692)
(Missing Cases)	(231)	(18)
LASTGPA(%)		
Mean	2.23	2.5
SD	0.9	0.8
(Valid Cases)	(3,982)	(681)
(Missing Cases)	(300)	(29)
TOTUNITS(%)		
Mean	12.27	13.05
SD	3.16	2.8
(Valid Cases)	(4,026)	(692)
(Missing Cases)	(256)	(18)

```
C     MIX.FOR
C
C     DIMENSIONS: NP =  8 = NUMBER X'S
C                 NY =  4 = NUMBER Y'S
C                 NVAR = 9 = NUMBER X'S AND Y'S
C                 NCELLS = 32  = NUMBER CELLS
C                 NOBS= 300 = NUMBER OBSERVATIONS
C                 7 = MAX. NUMBER CATEGORIES FOR A SINGLE Y
      IMPLICIT REAL*8(A-H,O-Z)
      DIMENSION IX(8),IY(4),NLEV(4),CODEM(4),
     1          XX(9),X(8,300),MDELBT(8,300),Y(4,300),W(32,300),
     2          PTAB(40,40),AMEAN(8,32),PI(32),XMEAN(8,32),XPI(32),
     3          SIGMA(8,8),PIVOT(40),UBAR(8),USQR(8,8),BPU(8),BPU2(8,8),
     4          BUB(8,8),REG(8),UU(40)
C.....UNIT 5 = CONTROL LANGUAGE, 10=INPUT DATA, 11=OUTPUT,
      OPEN(UNIT=5,FILE='MIX.CTL',STATUS='OLD')
      OPEN(UNIT=10,FILE='MIX.DAT',STATUS='OLD')
      OPEN(UNIT=11,FILE='MIX.OUT',STATUS='NEW')
      OPEN(UNIT=14,FILE='MIX.FIN',STATUS='NEW')
C.....READ INPUT PARAMETERS FROM UNIT 9
      READ(5,10) NOBS,NY,NP,NITER,IETYPE,XMCODE
C.....IETYPE=0 MEANS USE UNBIASED EST.OF COVARIANCES IN ITERATIONS
C           1 MEANS USE THE TRUE ML ESTIMATES (SEE SUB EMITER)
  10  FORMAT(I5/I5/I5/I5/I5/D5.0)
      READ(5,11) (IX(J),J=1,NP)
      READ(5,11) (IY(J),J=1,NY)
      READ(5,11) (NLEV(J),J=1,NY)
   11 FORMAT(10I5)
      NCELLS=1
      DO 20 I=1,NY
         NCELLS=NCELLS*NLEV(I)
  20  CONTINUE
C.....................................................
C  READ IN DATA AND DEFINE MISSING INDICATORS
C.....................................................
      CALL RDDATA(NOBS,NY,NP,NCELLS,XMCODE,NLEV,CODEM,IX,IY,
     *            X,MDELBT,Y,W)
      N1=15
      IF(NOBS.LT.15) N1=NOBS
      WRITE(11,*) 'X VARIABLES FOR FIRST ',N1,' CASES'
      CALL PRT(X,NP,N1)
      WRITE(11,*) 'MDELBT MATRIX FOR FIRST ',N1,' CASES'
      CALL PIT(MDELBT,NP,N1)
      WRITE(11,*) 'CATEGORICAL VARS FOR FIRST ',N1,' CASES'
      CALL PRT(Y,NY,N1)
C     WRITE(11,*) 'W MATRIX FOR THE MULTINOMIAL CELLS'
C     CALL PRT(W,NCELLS,N1)
C.....................................................
C  CALCULATE INITIAL ESTIMATES OF PARAMETERS
C.....................................................
      NDIM=NP+NCELLS
      CALL INTEST(NP,NCELLS,NDIM,NOBS,X,W,MDELBT,PTAB,AMEAN,PI)
      WRITE(11,*) 'AMEAN, ROWS=VARIABLES, COLS = CELLS'
      CALL PRINTR(AMEAN,NP,NCELLS)
C     WRITE(11,*) 'PTAB MATRIX'
```

```
C     CALL PRINTR(PTAB,NDIM,NDIM)
      WRITE(11,*) 'VECTOR PI:',(PI(I),I=1,NCELLS)
C
C   DO EM ITERATIONS
C
      CALL EM2(NITER,NP,NOBS,NDIM,NCELLS,X,MDELBT,W,PI,AMEAN,PTAB,
     1     IETYPE,XMEAN,XPI,SIGMA,PIVOT,UBAR,USQR,BPU,BPU2,BUB,REG,UU)
      STOP
      END
C
      SUBROUTINE PRT(X,IX,JX)
      IMPLICIT REAL*8(A-H,O-Z)
      DIMENSION X(IX,JX)
      DO 100 J=1,JX
         WRITE(11,22)(X(I,J),I=1,IX)
   22    FORMAT(1X,15F10.3)
  100 CONTINUE
      WRITE(11,33)
   33 FORMAT(//1X)
      RETURN
      END
C
      SUBROUTINE PRINTR(X,IX,JX)
      IMPLICIT REAL*8(A-H,O-Z)
      DIMENSION X(IX,JX)
      DO 100 I=1,IX
         WRITE(11,22)(X(I,J),J=1,JX)
   22    FORMAT(1X,15F10.3)
  100 CONTINUE
      WRITE(11,33)
   33 FORMAT(//1X)
      RETURN
      END
C
      SUBROUTINE PIT(L,IX,JX)
      DIMENSION L(IX,JX)
      DO 100 J=1,JX
         WRITE(11,22)(L(I,J),I=1,IX)
   22    FORMAT(1X,15I5)
  100 CONTINUE
      WRITE(11,33)
   33 FORMAT(//1X)
      RETURN
      END
C
      SUBROUTINE PRINTI(L,IX,JX)
      DIMENSION L(IX,JX)
      DO 100 I=1,IX
         WRITE(11,22)(L(I,J),J=1,JX)
   22    FORMAT(1X,15I5)
  100 CONTINUE
      WRITE(11,33)
   33 FORMAT(//1X)
      RETURN
      END
```

```
      SUBROUTINE DISCRM(NP,NCELLS,NDIM,PTAB,X,MISS,W,PI,NPRES,
     1                  NPOSS)
C
C     CALCULATES POSTERIOR PROBABILITIES OF BEING IN EACH CELL FOR
C                THE CURRENT CASE
C     INPUT:
C        NP, NCELLS,NDIM=NP+NCELLS
C        NPRES = NUMBER OF X'S PRESENT FOR THAT CASE
C        PTAB(I,J) = PIVOT TABLEAU MATRIX WITH PRESENT X'S PIVOTED IN
C                    I,J = 1,...,NP+NCELLS
C        X(I) = VECTOR OF X'S FOR THAT CASE I=1,...,NP
C        MISS(I) = 0 IF X(I) PRESENT, 1 IF X(I) MISSING
C        W(J) = CURRENT POSTERIOR PROB OF CELL J FOR THAT CASE
C        PI(J) = CURRENT PRIOR PROB FOR CELL J FOR THAT CASE
C     OUTPUT:
C        W(J) = UPDATED POSTERIOR PROB OF CELL J FOR THAT CASE
C        NPOSS = NUMBER OF NONZERO W(J) FOR THAT CASE
C
      IMPLICIT REAL*8(A-H,O-Z)
      DIMENSION X(NP),MISS(NP),W(NCELLS),PI(NCELLS),PTAB(NDIM,NDIM)
C*****************************
C     IF(X(2).EQ.108) THEN
C        WRITE(11,*) 'PI,PTAB(2,J),J=5,8) IN DISCRM'
C        WRITE(11,*) (PI(J),J=1,4),(PTAB(2,J),J=5,8)
C     ENDIF
C*****************************
      SUMP=0.
      NPOSS=0
      IF(NPRES.EQ.0) THEN
C.....NO X'S ARE PRESENT FOR THIS CASE
         DO 50 JCEL=1,NCELLS
            IF(W(JCEL).GT.0.) THEN
               NPOSS=NPOSS+1
               W(JCEL)=PI(JCEL)
               SUMP=SUMP+PI(JCEL)
               ELSE
                  W(JCEL)=10D6
            ENDIF
  50     CONTINUE
      GO TO 100
      ENDIF
C.....AT LEAST ONE X IS PRESENT
C.....SFACT WILL BE THE LARGEST EXPONENT ARGUMENT (OVER CELLS)
      SFACT=-10D-10
      DO 70 JCEL=1,NCELLS
         IF(W(JCEL).GT.0.) THEN
            NPOSS=NPOSS+1
            W(JCEL)=(PTAB(NP+JCEL,NP+JCEL)+(1./PI(JCEL)))/2.
            W(JCEL)=DLOG(PI(JCEL))+W(JCEL)
            DO 60 IVAR=1,NP
               IF(MISS(IVAR).EQ.0) W(JCEL)=W(JCEL)+
     *         PTAB(IVAR,NP+JCEL)*X(IVAR)
  60        CONTINUE
```

```
            IF(W(JCEL).GT.SFACT) SFACT=W(JCEL)
            ELSE
               W(JCEL)=10D6
         ENDIF
   70  CONTINUE
       DO 80 JCEL=1,NCELLS
          IF(W(JCEL).NE.10D6) THEN
C.....PREVENT UNDERFLOW
            EXPON=W(JCEL)-SFACT
            IF(EXPON.LE.-180.) EXPON=-180.
               W(JCEL)=DEXP(EXPON)
            SUMP=SUMP+W(JCEL)
          ENDIF
   80  CONTINUE
  100  CONTINUE
       DO 120 JCEL=1,NCELLS
          IF(W(JCEL).EQ.10D6) THEN
            W(JCEL)=0.
            ELSE
               W(JCEL)=W(JCEL)/SUMP
          ENDIF
  120  CONTINUE
       RETURN
       END
       SUBROUTINE DPIVOT(A,N,PIVOT,U,DET)
C
C      DOUBLE PRECISION PIVOT ROUTINE
C            PIVOT(A11):
C            A11 A12  ->   -INV(A11)    INV(A11)A12
C            A21 A22  ->    A21INV(A11) A22-A21INV(A11)A12
C
C            ADAPTED FROM SWP SUBROUTINE ON PG 71 OF STATISTICAL
C                    METHODS FOR DIGITAL COMPUTERS, VOL III.
C
C      INPUT: A(NXN) = MATRIX TO BE PIVOTED
C             N = ORDER OF A
C             PIVOT(J)=1,0,-1 FOR PIVOT, LEAVE ALONE, REVERSE PIVOT
C             U (NX1) SCRATCH VECTOR
C      OUTPUT: A = PIVOTED MATRIX
C             DET = LN OF ABS VALUE OF DET. OF PIVOTED SUBMATRIX
C                 = SUM OF LNS OF ABS VALUES OF DETS OF PIVOTED
C                   SUBMATRICES IF BOTH PIVOTS AND REVERSE PIVOTS USED
C
       IMPLICIT REAL*8(A-H,O-Z)
       DIMENSION PIVOT(N),U(N),A(N,N)
       DET=0.
       DO 100 K=1,N
          IF(PIVOT(K).NE.0) THEN
            DET=DET+DLOG(DABS(A(K,K)))
            FLAG=-1.*PIVOT(K)
            C=A(K,K)
            DO 1 I=1,K
               U(I)=A(I,K)
```

```
                A(I,K)=0.
    1         CONTINUE
              DO 2 I=K,N
                 U(I)=A(K,I)
                 A(K,I)=0.
    2         CONTINUE
              U(K)=FLAG
              DO 4 I=1,N
                 DO 3 J=I,N
                    A(I,J)=A(I,J)-U(I)*U(J)/C
                    A(J,I)=A(I,J)
    3            CONTINUE
    4         CONTINUE
           ENDIF
  100   CONTINUE
        RETURN
        END
        SUBROUTINE EM2(NITER,NP,NOBS,NDIM,NCELLS,X,MDELBT,W,PI,
       1            AMEAN,PTAB,IETYPE,XMEAN,XPI,SIGMA,PIVOT,UBAR,USQR,
       2            BPU,BPU2,BUB,REG,UU)
C
C       DOES EM ITERATIONS
C
C       INPUT:
C          NITER = MAX NUMBER ITERATIONS
C          NP,NCELLS,NDIM=NP+NCELLS
C          X(IVAR,IOBS),MDELBT(IVAR,IOBS)
C          W(JCEL,IOBS)
C          AMEAN(IVAR,JCEL)
C          PTAB(NDIM,NDIM)
C          PI(JCEL)
C          IETYPE = 0 IF UNBIASED ESTS OF COVARIANCES USED IN ITERATIONS
C                   1 IF TRUE ML ESTIMATES TO BE USED
C
        IMPLICIT REAL*8(A-H,O-Z)
        DIMENSION X(NP,NOBS),MDELBT(NP,NOBS),W(NCELLS,NOBS),
       1          AMEAN(NP,NCELLS),PTAB(NDIM,NDIM),PI(NCELLS),
       2          XMEAN(NP,NCELLS),XPI(NCELLS),SIGMA(NP,NP),
       3          PIVOT(NDIM),UBAR(NP),USQR(NP,NP),BPU(NP),
       4          BPU2(NP,NP),BUB(NP,NP),REG(NP),UU(NDIM),ICPAT(20)
C
C.....LOOP OVER ITERATIONS
C
        OLDLIK=10D10
        DO 1000 ITER=1,NITER
C.........CLEAR ACCUMULATORS XMEAN,XPI,SIGMA,
C          XLLIK=LOG LIKELIHOOD FOR X'S, YLLIK = LOG LIK FOR Y'S
           XLLIK=0.
           YLLIK=0.
           DO 5 ICEL=1,NCELLS
              PIVOT(ICEL+NP)=0.
              UU(ICEL+NP)=0.
              XPI(ICEL)=0.
              DO 4 IVAR=1,NP
```

```
                XMEAN(IVAR,ICEL)=0.
    4         CONTINUE
    5       CONTINUE
            DO 7 IVAR=1,NP
C.....INITIALIZE INDICATOR FOR CURRENT PIVOT PATTERN
C             ICPAT(I)=0 IF ITH VAR PIVOTED IN, 1 IF NOT
              ICPAT(IVAR)=1.
              DO 6 JVAR=1,IVAR
                 SIGMA(IVAR,JVAR)=0.
                 SIGMA(JVAR,IVAR)=0.
    6         CONTINUE
    7       CONTINUE
C.....CURDET BELOW WILL BE LN(DET(SIGMA)) FOR CURRENT CASE
            CURDET=0.
C
C.....LOOP OVER OBSERVATIONS
C
            DO 900 IOBS=1,NOBS
C
C.....PIVOT TO CURRENT PATTERN OF MISSING, IF DIFFERENT
C               FROM PREVIOUS CASE
C               NPRES=NUMBER PRESENT X'S
C               PIVOT(J) = 1,0,-1 FOR PIVOT, LEAVE ALONE, REVERSE PIVOT
              NPRES=0.
              NDIFF=0.
              DO 10 IVAR=1,NP
                 IF(MDELBT(IVAR,IOBS).EQ.0) NPRES=NPRES+1
                 KK=MDELBT(IVAR,IOBS)-ICPAT(IVAR)
                 IF(KK.EQ.-1) THEN
                    NDIFF=NDIFF+1
                    DO 11 JKL=1,NP
                       PIVOT(JKL)=0.
                       UU(JKL)=0.
   11                  CONTINUE
                       PIVOT(IVAR)=1.
                       ICPAT(IVAR)=MDELBT(IVAR,IOBS)
                       CALL DPIVOT(PTAB,NDIM,PIVOT,UU,DET)
                       CURDET=CURDET+DET
                 ENDIF
                 IF(KK.EQ.1) THEN
                    NDIFF=NDIFF+1
                    DO 12 JKL=1,NP
                       PIVOT(JKL)=0.
                       UU(JKL)=0.
   12                  CONTINUE
                       PIVOT(IVAR)=-1.
                       ICPAT(IVAR)=MDELBT(IVAR,IOBS)
                       CALL DPIVOT(PTAB,NDIM,PIVOT,UU,DET)
                       CURDET=CURDET+DET
                 ENDIF
                 PIVOT(IVAR)=0.
   10          CONTINUE
               XLLIK=XLLIK-CURDET/2.
```

```
C.....CALCULATE POSTERIOR PROBABILITIES
      CALL DISCRM(NP,NCELLS,NDIM,PTAB,X(1,IOBS),MDELBT(1,IOBS),
     *     W(1,IOBS),PI,NPRES,NPOSS)
C
C.....CALCULATE THE REST OF THE TERMS IN THE LOG LIKLIHOOD
C
      CALL LIKHOD(NP,NCELLS,NDIM,X(1,IOBS),MDELBT(1,IOBS),AMEAN,
     1 PTAB,W(1,IOBS),PI,NPRES,XXL,YYL)
      XLLIK=XLLIK+XXL
      YLLIK=YLLIK+YYL
C.....INCREMENT XPI
      DO 15 JCEL=1,NCELLS
         XPI(JCEL)=XPI(JCEL)+W(JCEL,IOBS)
  15  CONTINUE
C.....CALCULATE WEIGHTED POSTERIOR SUMS
         CALL WMEANS(NP,NCELLS,NDIM,X(1,IOBS),MDELBT(1,IOBS),
     1                W(1,IOBS),AMEAN,PTAB,UBAR,USQR,
     2                BPU,BPU2,BUB,REG)
C.....IMPUTATIONS AND E-STEP ARE NEXT
      DO 100 IVAR=1,NP
C.....FILL IN ESTIMATES OF MISSING X'S
         IF(MDELBT(IVAR,IOBS).EQ.1) X(IVAR,IOBS)=REG(IVAR)-
     1   BPU(IVAR)+UBAR(IVAR)
C.....E-STEP FOR CELL MEANS
         DO 30 JCEL=1,NCELLS
            IF(MDELBT(IVAR,IOBS).EQ.0) THEN
               XXXX=X(IVAR,IOBS)
               ELSE
                 XXXX=AMEAN(IVAR,JCEL)+REG(IVAR)
                 DO 28 KVAR=1,NP
                    XXXX=XXXX-PTAB(IVAR,KVAR)*
     1              (1.-MDELBT(KVAR,IOBS))*AMEAN(KVAR,JCEL)
  28             CONTINUE
            ENDIF
            XMEAN(IVAR,JCEL)=XMEAN(IVAR,JCEL)+W(JCEL,IOBS)*XXXX
  30      CONTINUE
C.....E-STEP FOR COVARIANCES
         DO 90 JVAR=1,IVAR
            IF(MDELBT(IVAR,IOBS).EQ.0.AND.
     1      MDELBT(JVAR,IOBS).EQ.0) THEN
C.....IVAR AND JVAR PRESENT
               SINC=X(IVAR,IOBS)*X(JVAR,IOBS)
            ELSE IF(MDELBT(IVAR,IOBS).EQ.0.AND.
     1      MDELBT(JVAR,IOBS).EQ.1) THEN
C.....IVAR PRESENT, JVAR MISSING
               SINC=X(IVAR,IOBS)*(UBAR(JVAR)+REG(JVAR)-
     1              BPU(JVAR))
            ELSE IF(MDELBT(IVAR,IOBS).EQ.1.AND.
     1      MDELBT(JVAR,IOBS).EQ.0) THEN
C.....IVAR MISSING, JVAR PRESENT
               SINC=X(JVAR,IOBS)*(UBAR(IVAR)+REG(IVAR)-
     1              BPU(IVAR))
            ELSE IF(MDELBT(IVAR,IOBS).EQ.1.AND.
```

```
     1           MDELBT(JVAR,IOBS).EQ.1) THEN
C.....IVAR AND JVAR MISSING
                 SINC=PTAB(IVAR,JVAR)+USQR(IVAR,JVAR)
     1               +UBAR(IVAR)*REG(JVAR)-BPU2(JVAR,IVAR)
     2               +UBAR(JVAR)*REG(IVAR)-BPU2(IVAR,JVAR)
     3               +REG(IVAR)*REG(JVAR)-REG(IVAR)*BPU(JVAR)
     4               -REG(JVAR)*BPU(IVAR)+BUB(IVAR,JVAR)
                ENDIF
                SIGMA(IVAR,JVAR)=SIGMA(IVAR,JVAR)+SINC
   90         CONTINUE
  100       CONTINUE
  900    CONTINUE
C.....END OF LOOP OVER OBSERVATIONS
C
C.....CALCULATE AND PRINT OUT -2*LOG LIKEHOOD FOR X'S, Y'S, ALL
C
      XLLIK=-2.*XLLIK
      YLLIK=-2.*YLLIK
      TLLIK=XLLIK+YLLIK
      WRITE(11,*) 'PRIOR TO ITERATION NUMBER: ',ITER
      WRITE(11,*) '     -2*LN(LIKELIHOOD) FOR X"S = ',XLLIK
      WRITE(11,*) '     -2*LN(LIKELIHOOD) FOR Y"S = ',YLLIK
      WRITE(11,*) '         -2*FULL LN LIKELIHOOD = ',TLLIK
      IF(ITER.EQ.NITER) THEN
         WRITE(11,*) 'X FOR 1ST 15 CASES AFTER ITERATION NO.',ITER
         CALL PRT(X,NP,15)
         WRITE(11,*) 'W FOR 1ST 15 CASES AFTER ITERATION NO. ',ITER
         CALL PRT(W,NCELLS,15)
      ENDIF
C
C.....UPDATE  AMEAN,PI,PTAB IN M STEP
C
       CALL MSTEP(NP,NCELLS,NOBS,NDIM,SIGMA,XPI,XMEAN,PTAB,AMEAN,PI,
     1            IETYPE)
C
       NABC=MOD(ITER,5)
       IF(NABC.EQ.0) THEN
          WRITE(11,*) 'SIGMA MATRIX AFTER ITERATION NUMBER ',ITER
          CALL PRINTR(SIGMA,NP,NP)
          WRITE(11,*) 'CELL MEANS, ROWS=VARS, COLS=CELLS'
          CALL PRINTR(AMEAN,NP,NCELLS)
          WRITE(11,*) 'CELL PROBABILITIES'
          CALL PRINTR(PI,1,NCELLS)
       ENDIF
C
C.....CHECK FOR CONVERGENCE (TO BE WRITTEN LATER)
C
       DELLIK=OLDLIK-TLLIK
       IF(DELLIK.LT.0.00000001.AND.ITER.GT.3) GO TO 1001
       OLDLIK=TLLIK
 1000  CONTINUE
 1001  CONTINUE
C
C.....WRITE FINAL ESTIMATES TO FILE AND TO OUTPUT
```

```
C
      DO 1020 JCEL=1,NCELLS
C        WRITE(14,1021) PI(JCEL),(AMEAN(K,JCEL),K=1,NP)
         WRITE(11,1021) PI(JCEL),(AMEAN(K,JCEL),K=1,NP)
 1021    FORMAT(10F10.5)
 1020 CONTINUE
      DO 1030 IVAR=1,NP
C        WRITE(14,1021) (PTAB(IVAR,J),J=1,IVAR)
         WRITE(11,1021) (PTAB(IVAR,J),J=1,IVAR)
 1030 CONTINUE
C     DO 1040 IOBS=1,NOBS
C        WRITE(14,1041) (X(IVAR,IOBS),IVAR=1,NP),
C    *                  (W(JCEL,IOBS),JCEL=1,NCELLS)
C1041    FORMAT(11F8.3)
C1040 CONTINUE
      RETURN
      END
      SUBROUTINE INTEST(NP,NCELLS,NDIM,NOBS,X,W,MDELBT,PTAB,AMEAN,PI)
C
C     CALCULATES INITIAL ESTIMATES OF PARAMETERS
C                SET UP TO USE CASES WITH COMPLETE X DATA
C                FOR MEANS AND COVARIANCES, ALL DATA FOR PI(J)
C                UNPOOLED COV. MATRIX IS USED AS INITIAL ESTIMATE
C
C     INPUT:
C        NP,NCELLS,NDIM=NP+NCELLS,NOBS
C        X(IVAR,IOBS)
C        MDELBT(IVAR,IOBS): (0 FOR PRESENT, 1 FOR MISSING)
C        W(JCEL,IOBS): INITIAL EST. OF POST. PROB FOR CELL JCEL
C     READ IN FROM CONTROL FILE (UNIT 9):
C        MODE: 0=COMPUTE INIT. ESTIMATES,
C              1=READ IN FROM UNIT 13 = /USR/MDS/EM.INITIAL
C     OUTPUT:
C        AMEAN(IVAR,JCEL): MEAN OF X(IVAR) FOR CELL JCEL
C        PTAB(I,J): PIVOT TABLEAU MATRIX
C                   COMPL. CASE COVARIANCE MATRIX (UNPOOLED)
C                   FOR I,J LE NP
C                   CELL MEANS FOR I LE NP, J=NP+1,...,NP+NCELLS
C                   DIAG MATRIX OF -1/CELL PROB FOR I,J GT NP
C        PI(JCEL): PROB(CELL JCEL) IN MULTINOMIAL MODEL
C     OTHER:
C        WT(J): COUNT OF NUMBER COMPLETE CASES IN CELL J
C        XBAR(I): MEAN OF X(I) ACROSS CELLS FOR COMPLETE OBSERVATIONS
      IMPLICIT REAL*8(A-H,O-Z)
      DIMENSION X(NP,NOBS),MDELBT(NP,NOBS),W(NCELLS,NOBS),
     *          AMEAN(NP,NCELLS),PI(NCELLS),
     *          PTAB(NDIM,NDIM),WT(50),XBAR(5),NCMPL(4,12)
C
C.....READ MODE OF CALCULATING INITIAL ESTIMATES
C
      READ(5,1) MODE
   1  FORMAT(I5)
      IF(MODE.EQ.1) THEN
         DO 3 JCEL=1,NCELLS
```

```
            READ(5,2) PI(JCEL),(AMEAN(K,JCEL),K=1,NP)
    2       FORMAT(10D10.5)
    3     CONTINUE
          DO 5 IVAR=1,NP
             READ(5,2) (PTAB(IVAR,J),J=1,IVAR)
             DO 4 JVAR=1,IVAR
                PTAB(JVAR,IVAR)=PTAB(IVAR,JVAR)
    4        CONTINUE
    5     CONTINUE
          GO TO 301
       ENDIF
C.....INITIALIZE COUNTERS
       DO 10 JCEL=1,NCELLS
         PI(JCEL)=0.
         WT(JCEL)=0.
         DO 6 IVAR=1,NP
            AMEAN(IVAR,JCEL)=0.
    6    CONTINUE
   10  CONTINUE
       DO 20 IVAR=1,NP
          XBAR(IVAR)=0.
          DO 15 JVAR=1,IVAR
             PTAB(I,J)=0.
             PTAB(J,I)=0.
   15     CONTINUE
   20  CONTINUE
       NCOBS=0
C.....LOOP OVER CASES
       DO 100 IOBS=1,NOBS
C........CUMULATE PI(J)=PR(CELL J)
          DO 25 J=1,NCELLS
             PI(J)=PI(J)+W(J,IOBS)
   25     CONTINUE
C........SKIP THE REST FOR THIS CASE IF ANY X'S ARE MISSING
          DO 30 IVAR=1,NP
             IF(MDELBT(IVAR,IOBS).EQ.1) GO TO 100
   30     CONTINUE
          NCOBS=NCOBS+1
C........CALCULATE AMEANS
          DO 40 IVAR=1,NP
             XBAR(IVAR)=XBAR(IVAR)+X(IVAR,IOBS)
             DO 35 JCEL=1,NCELLS
                 AMEAN(IVAR,JCEL)=AMEAN(IVAR,JCEL)+
     *          X(IVAR,IOBS)*W(JCEL,IOBS)
   35        CONTINUE
   40     CONTINUE
          DO 45 JCEL=1,NCELLS
             WT(JCEL)=WT(JCEL)+W(JCEL,IOBS)
   45     CONTINUE
  100  CONTINUE
C.....DIVIDE MEANS BY SAMPLE SIZES
       DO 110 IVAR=1,NP
          XBAR(IVAR)=XBAR(IVAR)/NCOBS
```

```
          DO 105 JCEL=1,NCELLS
             IF(WT(JCEL).GT.0.) THEN
                AMEAN(IVAR,JCEL)=AMEAN(IVAR,JCEL)/WT(JCEL)
                ELSE
                   AMEAN(IVAR,JCEL)=XBAR(IVAR)
                   IF(IVAR.EQ.1) THEN
                    WRITE(11,*) 'NO COMPLETE CASES IN CELL ',JCEL
                    WRITE(11,*) 'MEANS SET EQUAL TO RAW MEANS'
                   ENDIF
             ENDIF
 105      CONTINUE
 110   CONTINUE
C.....NOW CALCUATE UNPOOLED COVARIANCES FROM COMPLETE CASES
       DO 200 IOBS=1,NOBS
          DO 160 IVAR=1,NP
             IF(MDELBT(IVAR,IOBS).EQ.1) GO TO 200
             TIVAR=X(IVAR,IOBS)-XBAR(IVAR)
             DO 150 JVAR=1,IVAR
                IF(MDELBT(JVAR,IOBS).EQ.1) GO TO 200
                PTAB(IVAR,JVAR)=PTAB(IVAR,JVAR)+TIVAR*
     *          (X(JVAR,IOBS)-XBAR(JVAR))
 150         CONTINUE
 160      CONTINUE
 200   CONTINUE
       DO 300 IVAR=1,NP
          DO 250 JVAR=1,IVAR
             PTAB(IVAR,JVAR)=PTAB(IVAR,JVAR)/(NCOBS-1.)
             PTAB(JVAR,IVAR)=PTAB(IVAR,JVAR)
 250      CONTINUE
 300   CONTINUE
C*************************************************
       DO 315 JCEL=1,NCELLS
          DO 314 IVAR=1,NP
             AMEAN(IVAR,JCEL)=0.
             NCMPL(IVAR,JCEL)=0
             DO 313 IOBS=1,NOBS
                IF(MDELBT(IVAR,IOBS).EQ.0.AND.W(JCEL,IOBS).EQ.1.) THEN
                   AMEAN(IVAR,JCEL)=AMEAN(IVAR,JCEL)+X(IVAR,IOBS)
                   NCMPL(IVAR,JCEL)=NCMPL(IVAR,JCEL)+1
                ENDIF
 313         CONTINUE
          IF(NCMPL(IVAR,JCEL).GT.0) THEN
             AMEAN(IVAR,JCEL)=AMEAN(IVAR,JCEL)/NCMPL(IVAR,JCEL)
             ELSE
                AMEAN(IVAR,JCEL)=XBAR(IVAR)
          ENDIF
 314      CONTINUE
 315   CONTINUE
C*************************************************
 301   CONTINUE
C.....FILL IN PIVOT TABLEAU WITH INITIAL ESTIMATES
       DO 350 IVAR=1,NP
          DO 320 JCEL=1,NCELLS
```

```
            PTAB(IVAR,NP+JCEL)=AMEAN(IVAR,JCEL)
            PTAB(NP+JCEL,IVAR)=AMEAN(IVAR,JCEL)
  320     CONTINUE
  350   CONTINUE
        DO 400 JCEL=1,NCELLS
           IF(MODE.EQ.0) PI(JCEL)=PI(JCEL)/NOBS
           DO 380 KCEL=1,JCEL
              PTAB(NP+JCEL,NP+KCEL)=0.
              IF(JCEL.EQ.KCEL) PTAB(NP+JCEL,NP+KCEL)=-1./PI(JCEL)
              PTAB(NP+KCEL,NP+JCEL)=PTAB(NP+JCEL,NP+KCEL)
  380     CONTINUE
  400   CONTINUE
        RETURN
        END
        SUBROUTINE LIKHOD(NP,NCELLS,NDIM,X,MISS,AMEAN,PTAB,
       1                  W,PI,NPRES,XXL,YYL)
C
C     CALCULATES CONTRIBUTION TO LIKELIHOOD FOR CURRENT CASE
C                (EXCEPT FOR THE TERM INVOLVING THE DETERMINANT
C                 OF THE COVARIANCE MATRIX)
C
C     INPUT: NP,NCELLS,NDIM
C            X(I) = VECTOR OF X'S FOR CURRENT CASE
C            MISS(I) = 0 IF X(I) PRESENT, 1 IF MISSING
C            W(J) = POST. PROB FOR CELL J, J=1,...,NCELLS
C            PI(J) = CURRENT EST. OF PROB. FOR CELL J
C            AMEAN(I,J) = MEAN OF X(I) IN CELL J
C            PTAB(I,J) = PIVOT TABLEAU MATRIX
C            NPRES = NUMBER OF NONMISSING X'S FOR THAT CASE
C     OUTPUT:
C            XXL = LN OF SUM OVER J: W(J)*EXP(X-MU)'(SIGMA INV)(X-MU))
C                = ZERO IF ALL X'S ARE MISSING (NPRES=0)
C            YYL = LN OF PROB. OF CELL IN COLLAPSED TABLE WHERE THE
C                  CURRENT OBS IS KNOWN TO BE
C                = ZERO IF ALL CATEG. VARS ARE MISSING (NPOSS=NCELLS))
C                  (BECAUSE THEN YYL=ALOG(SPROB)=ALOG(1)=0)
C
        IMPLICIT REAL*8(A-H,O-Z)
        DIMENSION X(NP),MISS(NP),AMEAN(NP,NCELLS),W(NCELLS),
       1          PTAB(NDIM,NDIM),PI(NCELLS)
        DATA AL2PI/1.837877D0/
C.....AL2PI=ALOG(2*3.1416)
        XXL=0.
        YYL=0.
        SPROB=0.
        DO 100 JCEL=1,NCELLS
           IF(W(JCEL).GT.0.) THEN
              SPROB=SPROB+PI(JCEL)
              IF(NPRES.EQ.0) GO TO 100
              SUM=0.
              DO 90 IVAR=1,NP
                 TI=(X(IVAR)-AMEAN(IVAR,JCEL))*(1.-MISS(IVAR))
                 DO 80 JVAR=1,IVAR
```

```
                  TJ=(X(JVAR)-AMEAN(JVAR,JCEL))*(1.-MISS(JVAR))
                  IF(IVAR.EQ.JVAR) THEN
                     SUM=SUM+TI*TJ*PTAB(IVAR,JVAR)
                     ELSE
                        SUM=SUM+2.*TI*TJ*PTAB(IVAR,JVAR)
                  ENDIF
   80          CONTINUE
   90       CONTINUE
C...........PTAB ABOVE WAS NEGATIVE OF INVERSE, SO SUM IS NEGATIVE
C...........PREVENT UNDERFLOW
            IF(SUM.LT.-360.) SUM=-360.
            XXL=XXL+PI(JCEL)*DEXP(0.5*SUM)
         ENDIF
  100  CONTINUE
       IF(NPRES.GT.0) XXL=DLOG(XXL/SPROB)-(NPRES/2.)*AL2PI
       YYL=DLOG(SPROB)
       RETURN
       END
       SUBROUTINE MSTEP(NP,NCELLS,NOBS,NDIM,SIGMA,XPI,XMEAN,PTAB,
      *                 AMEAN,PI,IETYPE)
C
C      INPUT:
C        NP,NCELLS,NR,NOBS,NDIM,NRP=NR+NP
C        NR=4 HERE, NP=4
C        ADES(NR X NCELLS)
C        SIGMA(NP X NP)
C        XMEAN(NP X NCELLS)
C        XPI(NCELLS)
C        PI(NCELLS)
C        PTAB(NDIM X NDIM)
C        ADA(NR X NR)
C        BPARM(NP X NR)
C        PMST(NR+NP X NR+NP) = MATRIX PIVOTED IN M-STEP
C
       IMPLICIT REAL*8(A-H,O-Z)
       DIMENSION ADES(6,12),SIGMA(NP,NP),XMEAN(NP,NCELLS),BPARM(4,6),
      *          AMEAN(NP,NCELLS),PI(NCELLS),XPI(NCELLS),
      *          PTAB(NDIM,NDIM),PMST(10,10),U(50),PIVOT(50),AAAA(12,6)
C      DATA ADES/1.,1.,0.,0.,1.,0.,1.,0.,1.,0.,0.,1.,1.,-1.,-1.,-1.,
C     1          1.,1.,0.,0.,1.,0.,1.,0.,1.,0.,0.,1.,1.,-1.,-1.,-1.,
C     2          1.,1.,0.,0.,1.,0.,1.,0.,1.,0.,0.,1.,1.,-1.,-1.,-1./
       DATA AAAA/1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.,
      1          1., 1.,-1.,-1., 1., 1.,-1.,-1., 1., 1.,-1.,-1.,
      2          1.,-1., 1.,-1., 1.,-1., 1.,-1., 1.,-1., 1.,-1.,
      3          1.,-1.,-1., 1., 1.,-1.,-1., 1., 1.,-1.,-1., 1.,
      4          1., 1., 1., 1., 0., 0., 0., 0.,-1.,-1.,-1.,-1.,
      5          0., 0., 0., 0., 1., 1., 1., 1.,-1.,-1.,-1.,-1./
C     6          1., 1.,-1.,-1., 0., 0., 0., 0.,-1.,-1., 1., 1.,
C     7          1.,-1.. 1.,-1., 0., 0., 0., 0.,-1., 1.,-1., 1.,
C     8          0., 0., 0., 0., 1., 1.,-1.,-1.,-1.,-1., 1., 1.,
C     9          0., 0., 0., 0., 1.,-1., 1.,-1.,-1., 1.,-1., 1./
C
       DO 1 I=1,12
```

```
         DO 1 J=1,6
            ADES(J,I)=AAAA(I,J)
    1  CONTINUE
C      CALL PRINTR(ADES,6,12)
C
       NR=NCELLS
       NRP=NR+NP
C
C      DO THE FOLLOWING IF NO RESTRICTIONS ON MEANS WANTED (NR=NCELLS)
C
       IF(NR.EQ.NCELLS) THEN
          DO 6 I=1,NDIM
             DO 5 J=1,I
                PTAB(I,J)=0.
                PTAB(J,I)=0.
    5        CONTINUE
    6     CONTINUE
          DO 15 JCEL=1,NCELLS
             DO 12 IVAR=1,NP
                XMEAN(IVAR,JCEL)=XMEAN(IVAR,JCEL)/XPI(JCEL)
                AMEAN(IVAR,JCEL)=XMEAN(IVAR,JCEL)
                PTAB(IVAR,NP+JCEL)=AMEAN(IVAR,JCEL)
                 PTAB(NP+JCEL,IVAR)=AMEAN(IVAR,JCEL)
                DO 10 JVAR=1,IVAR
                   PTAB(IVAR,JVAR)=PTAB(IVAR,JVAR)-
     1             XPI(JCEL)*XMEAN(IVAR,JCEL)*XMEAN(JVAR,JCEL)
   10           CONTINUE
   12        CONTINUE
   15     CONTINUE
          DDD1=NOBS-NCELLS
          DDD2=NOBS
          DO 30 I=1,NP
             DO 25 J=1,I
             PTAB(I,J)=(PTAB(I,J)+SIGMA(I,J))/DDD1
             IF(IETYPE.EQ.1) PTAB(I,J)=PTAB(I,J)*(DDD1/DDD2)
             PTAB(J,I)=PTAB(I,J)
             SIGMA(I,J)=PTAB(I,J)
             SIGMA(J,I)=SIGMA(I,J)
   25        CONTINUE
   30     CONTINUE
          ELSE
C
          DO 40 I=1,NR
             DO 38 J=1,I
                PMST(I,J)=0.
                DO 37 K=1,NCELLS
                   PMST(I,J)=PMST(I,J)+ADES(I,K)*ADES(J,K)*XPI(K)
   37           CONTINUE
             PMST(J,I)=PMST(I,J)
   38        CONTINUE
   40     CONTINUE
          DO 50 I=1,NR
             DO 49 J=1,NP
```

```
            PMST(I,J+NR)=0.
            DO 48 K=1,NCELLS
               PMST(I,J+NR)=PMST(I,J+NR)+ADES(I,K)*XMEAN(J,K)
 48         CONTINUE
            PMST(J+NR,I)=PMST(I,J+NR)
 49      CONTINUE
 50   CONTINUE
      DO 60 I=1,NP
         DO 59 J=1,I
            PMST(I+NR,J+NR)=SIGMA(I,J)
            PMST(J+NR,I+NR)=PMST(I+NR,J+NR)
 59      CONTINUE
 60   CONTINUE
      DO 70 I=1,NRP
         PIVOT(I)=1.
         IF(I.GT.NR) PIVOT(I)=0.
         U(I)=0.
 70   CONTINUE
      CALL DPIVOT(PMST,NRP,PIVOT,U,DET)
C.....PARAMETERS IN BPARM AND NEW ESTIMATES OF CELL MEANS:
      DO 80 I=1,NR
         DO 75 J=1,NP
            BPARM(J,I)=PMST(I,J+NR)
 75      CONTINUE
 80   CONTINUE
      WRITE(11,*) 'BPARM MATRIX'
      CALL PRINTR(BPARM,NP,NR)
      DO 90 I=1,NP
         DO 89 J=1,NCELLS
            AMEAN(I,J)=0.
            DO 88 K=1,NR
               AMEAN(I,J)=AMEAN(I,J)+BPARM(I,K)*ADES(K,J)
 88         CONTINUE
            PTAB(I,J+NP)=AMEAN(I,J)
            PTAB(J+NP,I)=AMEAN(I,J)
 89      CONTINUE
 90   CONTINUE
      DDD1=NOBS-NR
      DDD2=NOBS
      DO 100 I=1,NP
         DO 95 J=1,I
            SIGMA(I,J)=PMST(I+NR,J+NR)/DDD1
            IF(IETYPE.EQ.1) SIGMA(I,J)=SIGMA(I,J)*(DDD1/DDD2)
            PTAB(I,J)=SIGMA(I,J)
            PTAB(J,I)=SIGMA(I,J)
            SIGMA(J,I)=SIGMA(I,J)
 95      CONTINUE
100   CONTINUE
      ENDIF
C
C     NOW DO M-STEP FOR CELL PROBABILITIES
C
      DO 105 JCEL=1,NCELLS
```

```
         PI(JCEL)=XPI(JCEL)/NOBS
 105  CONTINUE
C*************************************************
C       SPECIAL M-STEP FOR ST LOUIS EXAMPLE
C
       GO TO 926
 922   CONTINUE
       PRSK0=PI(1)+PI(2)+PI(3)+PI(4)
       PRSK1=PI(5)+PI(6)+PI(7)+PI(8)
       PRSK2=PI(9)+PI(10)+PI(11)+PI(12)
       PLL=PI(1)+PI(5)+PI(9)
       PLH=PI(2)+PI(6)+PI(10)
       PHL=PI(3)+PI(7)+PI(11)
       PHH=PI(4)+PI(8)+PI(12)
       PI(1)=PRSK0*PLL
       PI(2)=PRSK0*PLH
       PI(3)=PRSK0*PHL
       PI(4)=PRSK0*PHH
       PI(5)=PRSK1*PLL
       PI(6)=PRSK1*PLH
       PI(7)=PRSK1*PHL
       PI(8)=PRSK1*PHH
       PI(9)=PRSK2*PLL
       PI(10)=PRSK2*PLH
       PI(11)=PRSK2*PHL
       PI(12)=PRSK2*PHH
 926   CONTINUE
C*************************************************
      DO 120 I=1,NCELLS
         DO 115 J=1,NCELLS
            PTAB(I+NP,J+NP)=0.
            IF(I.EQ.J) PTAB(I+NP,J+NP)=-1./PI(I)
 115     CONTINUE
 120  CONTINUE
      RETURN
      END
      SUBROUTINE RDDATA(NOBS,NY,NP,NCELLS,XMCODE,NLEV,CODEM,IX,IY,
     *             X,MDELBT,Y,W)
C     INPUT:
C        NOBS,NY,NP,NCELLS,
C        XMCODE = MISSING VALUE CODE FOR X'S
C        NLEV(I) = NUMBER OF CATEGORIES FOR Y(I), I=1,....NY
C        IX(I) = SEQUENCE NUMBER OF X(I) IN INPUT DATA
C        IY(I) = SEQUENCE NUMBER OF Y(I) IN INPUT DATA
C     READS FROM UNIT 9:
C        FMT = INPUT FORMAT
C        CODE(I,J) = CODE FOR JTH CATEGORY, ITH VARIBLE J=1,...,NLEV(I)
C        CODEM(I) = MISSING VALUE CODE FOR ITH CATEGOR. VAR
C        Y(I,J) = ITH CAT. VAR FOR JTH OBS, J=1,...,NOBS, I=1,...,NY
C        X(K,J) = KTH CONTIN. VAR FOR JTH OBS, J=1,...,NOBS K=1,...,NP
C        IX(J) = SEQUENCE NUMBER OF X(J) IN INPUT VARIABLES
C        IY(J) = SEQUENCE NUMBER OF Y(J) IN INPUT VARIABLES
C     OTHER VARIABLES:
```

```
C        MDELBT(K,J) = 0 IF X(K,J) PRESENT, 1 IF IT IS MISSING
C        NCELLS = NLEV(1)*NLEV(2)*...*NLEV(NY)
C        CAT(I,J) = 1/(NLEV(I)) IF Y(I) MISSING, J=1,...,NLEV(I)
C                   1           IF Y(I) = CODE(J)
C                   0           IF Y(I) NE CODE(J) AND NOT MISSING
C        W(L,J) =  INITIAL ESTIMATE OF CELL L FOR OBSERVATION J
C                  L=1,...,NCELLS,  J=1,...,NOBS
C
      IMPLICIT REAL*8(A-H,O-Z)
      DIMENSION IX(NP),IY(NY),NLEV(NY),CODEM(NY),CODE(4,7),XX(9),
     *X(NP,NOBS),CAT(4,7),MDELBT(NP,NOBS),Y(NY,NOBS),W(NCELLS,NOBS)
      DIMENSION PROD(4)
      CHARACTER*4 FMT(18)
C
      READ(5,2) FMT
    2 FORMAT(18A4)
      READ(5,13) (CODEM(J),J=1,NY)
   13 FORMAT(10D5.0)
      DO 20 I=1,NY
         READ(5,16) (CODE(I,J),J=1,NLEV(I))
   16    FORMAT(10D5.0)
   20 CONTINUE
C.....READ IN DATA
      NVARS=NP+NY
      DO 200 IOBS=1,NOBS
         IF(NY.EQ.1.AND.NLEV(1).EQ.1) THEN
            READ(10,FMT) (XX(K),K=1,NVARS-1)
            ELSE
               READ(10,FMT) (XX(K),K=1,NVARS)
         ENDIF
         DO 40 K=1,NP
            X(K,IOBS)=XX(IX(K))
            MDELBT(K,IOBS)=0
            IF(X(K,IOBS).EQ.XMCODE) MDELBT(K,IOBS)=1
   40    CONTINUE
C........IF NO CATEGORICAL VARIABLES (NY=1, NLEV(1)=1)
         IF(NY.EQ.1.AND.NLEV(1).EQ.1) THEN
            Y(1,IOBS)=1.
            W(1,IOBS)=1.
            K=1.
            GO TO 200
         ENDIF
         DO 50 I=1,NY
            Y(I,IOBS)=XX(IY(I))
            DO 45 J=1,NLEV(I)
               CAT(I,J)=0.
               IF(Y(I,IOBS).EQ.CODE(I,J)) CAT(I,J)=1.
               IF(Y(I,IOBS).EQ.CODEM(I)) CAT(I,J)=1./FLOAT(NLEV(I))
   45       CONTINUE
   50    CONTINUE
C........FIND W(L,IOBS)
         K=0
         DO 100 I1=1,NLEV(1)
```

```
            IF(NY.EQ.1) THEN
               K=K+1
               W(K,IOBS)=CAT(1,I1)
               GO TO 100
            ENDIF
            PROD(1)=CAT(1,I1)
            DO 90 I2=1,NLEV(2)
               IF(NY.EQ.2) THEN
                  K=K+1
                  W(K,IOBS)=PROD(1)*CAT(2,I2)
                  GO TO 90
               ENDIF
               PROD(2)=PROD(1)*CAT(2,I2)
               DO 80 I3=1,NLEV(3)
                  IF(NY.EQ.3) THEN
                     K=K+1
                     W(K,IOBS)=PROD(2)*CAT(3,I3)
                     GO TO 80
                  ENDIF
                  PROD(3)=PROD(2)*CAT(3,I3)
                  DO 70 I4=1,NLEV(4)
                     IF(NY.EQ.4) THEN
                        K=K+1
                        W(K,IOBS)=PROD(3)*CAT(4,I4)
                        GO TO 70
                     ENDIF
                     PROD(4)=PROD(3)*CAT(4,I4)
                     DO 60 I5=1,NLEV(5)
                        IF(NY.EQ.5) THEN
                           K=K+1
                           W(K,IOBS)=PROD(4)*CAT(5,I5)
                           GO TO 60
                        ENDIF
   60                CONTINUE
   70             CONTINUE
   80          CONTINUE
   90       CONTINUE
  100    CONTINUE
  200 CONTINUE
      NCELLS=K
      RETURN
      END
      SUBROUTINE WMEANS(NP,NCELLS,NDIM,X,M,W,AMEAN,PTAB,UBAR,USQR,
     1                  BPU,BPU2,BUB,REG)
C
C     CALCULATES QUANTITIES USED IN IMPUTATIONS AND E-STEP
C
C     INPUT:
C        NP,NCELLS,NDIM
C        AMEAN(IVAR,JCEL) = MEAN OF X(IVAR) IN CELL JCEL
C        PTAB(I,J)
C        W(K) = POSTERIOR PROB. OF CELL K
C        X(I),I=1,...,NP = VECTOR OF CONTINUOUS DATA FOR THAT CASE
```

```
C        M(I),I=1,...,NP = 0 IF X(I) PRESENT, 1 IF MISSING
C     OUTPUT:
C        UBAR(I) = SUM OVER J=1 TO NP OF W(J)*AMEAN(I,J)
C        USQR(I,K) = SUM OVER J=1 TO NP OF W(J)*AMEAN(I,J)*AMEAN(K,J)
C        BPU(I)      SEE DOCUMENTATION
C        BPU2(I,K)   SEE DOCUMENTATION
C        BUB(I,K)    SEE DOCUMENTATION
C        REG(I)      SEE DOCUMENTATION
C
      IMPLICIT REAL*8(A-H,O-Z)
      DIMENSION AMEAN(NP,NCELLS),UBAR(NP),USQR(NP,NP),W(NCELLS),
     1          X(NP),M(NP),PTAB(NDIM,NDIM),BPU(NP),BPU2(NP,NP),
     2          REG(NP),BUB(NP,NP)
      DO 10 I=1,NP
         UBAR(I)=0.
         DO 5 J=1,NCELLS
            UBAR(I)=UBAR(I)+W(J)*AMEAN(I,J)
    5    CONTINUE
   10 CONTINUE
      DO 20 I=1,NP
         DO 15 J=1,I
            USQR(I,J)=0.
            DO 12 K=1,NCELLS
               USQR(I,J)=USQR(I,J)+W(K)*AMEAN(I,K)*AMEAN(J,K)
   12       CONTINUE
            USQR(J,I)=USQR(I,J)
   15    CONTINUE
   20 CONTINUE
C
C ....CALCULATE BUP(I) AND REG(I)
C
      DO 30 I=1,NP
         BPU(I)=0.
         REG(I)=0.
         IF(M(I).EQ.1) THEN
            DO 25 J=1,NP
               BETA=PTAB(I,J)
               BPU(I)=BPU(I)+(1.-M(J))*BETA*UBAR(J)
               REG(I)=REG(I)+(1.-M(J))*BETA*X(J)
   25       CONTINUE
         ENDIF
   30 CONTINUE
C
C.....CALCULATE BPU2(I,J)
C
      DO 50 I=1,NP
         DO 45 J=1,NP
            BPU2(I,J)=0.
            IF(M(1).EQ.1) THEN
               DO 40 K=1,NP
                  BETA=PTAB(I,K)
                  BPU2(I,J)=BPU2(I,J)+(1.-M(K))*BETA*USQR(J,K)
   40          CONTINUE
```

```
            ENDIF
   45     CONTINUE
   50  CONTINUE
C
C.....CALCULATE BUB(I,J)
C
       DO 80 I=1,NP
          DO 75 J=1,I
             BUB(I,J)=0.
             IF(M(I).EQ.1.AND.M(J).EQ.1) THEN
                DO 70 K1=1,NP
                   DO 60 K2=1,NP
                      BETA1=PTAB(I,K1)
                      BETA2=PTAB(J,K2)
                      BUB(I,J)=BUB(I,J)+(1.-M(K1))*(1.-M(K2))*BETA1*
     1                            BETA2*USQR(K1,K2)
   60                CONTINUE
   70             CONTINUE
             ENDIF
             BUB(J,I)=BUB(I,J)
   75     CONTINUE
   80  CONTINUE
       RETURN
       END
```

References

Afifi, A. A., & Elashoff, R. M. (1969). Missing observations in multivariate statistics IV. A note on simple linear regression. Journal of the American Statistical Association, 64, 359-365.

Afifi, A. A., & Elashoff, R. M. (1967). Missing observaitons in multivariate statistics II. Point estimation in simple linear regression. Journal of the American Statistical Association, 62, 10-29.

Afifi, A. A., & Elashoff, R. M. (1969). Missing observaitons in multivariate statistics III. Large sample analysis of simple linear regression. Journal of the American Statistical Association, 64, 337-358.

Afifi, A. A., & Elashoff, R. M. (1969). Missing observations in multivariate statistics IV. A note on simple linear regression. Journal of the American Statistical Association, 64, 359-365.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. Journal of the American Statistical Association, 52, 200-203.

Bailar, B. A., & Bailar, J. C. (1983). Comparison of the biases of the "hot deck" imputation procedure with an "equal weights" imputation procedure. In W. G. Madow & I. Olkin (Eds.), Incomplete Data in Sample Surveys, Vol. III: Symposium on Incomplete Data Proceedings. New York: Academic Press.

Baum, L. E., Petrie, T., Soules, G. & Weiss, N. (1970). A maximizaition technique occurring in the statistical analysis of probablistic functions of Markov chains. The Annals of Mathematical Statistics, 41(1), 164-171.

Beale, E. M., & Little, R. J. A. (1974). Missing values in multivariate analysis. Journal of Roval Statistical Society (Series B), 22, 302-306.

Beale, E. M., & Little R. J. A. (1975). Missing values in multivariate analysis. Journal of Roval Statistical Society (Series B), 22, 129-145.

Berk, R. A. (1983). An introduction to sample selection bias in sociological data. American Sociological Review, 48, 386-398.

Bhargava, R. (1962). Multivariate tests of hypotheses with incomplete data. Applied Mathematics and Statistical Laboratories Technical Report, 3.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. Journal of Royal Statistical Society (Series B), 22, 302-306.

Chen, T., & Fienberg, S. E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. Biometrics, 30, 629-642.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, N.J.: L. Erlbarum Associates.

Dempster, A. P., & Rubin, D. B. (1983). Overview, in W. G. Madow, I. Olkin, & D. B. Rubin (Eds.). Incomplete Data in Sample Surveys: Vol. II. Theory and Annotated Bibliography. New York: Academic Press.

Dempster, A., & Laird, N., & Rubin, D. B. (1982). Maximum likelihood from incomplete data via the EM algorithm. Journal of Royal Statistical Society (Series B), 39, 1-39.

Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. Journal of American Statistical Association, 77, 270-278.

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

Hartley, H. O. (1958). Maximum likelihood estimation from incomplete data. Biometrics, 14, 174-194.

Hartley, H. O., & Hocking R. R. (1971). The analysis of incomplete data. Biometrics, 27, 783-823.

Kalton, G., & Kasprzyk, D. (1982). Imputing for missing survey responses, Proceedings of the Survey Research Methods (pp. 22-31). Washington, D.C.: American Statistical Association.

Kim, J., & Curry, J. (1977). The treatment of missing data in multivariate analysis. Sociological Methods and Research, 6(2), 215-240.

Krzanowski, W. J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. Biometrics, 36, 493-499.

Krzanowski, W. J. (1982). Mixtures of continuous and categorical variables in discriminant analysis: A hypothesis-testing approach. Biometrics, 38, 991-1002.

Lavin, D. E., Alba, R. D., & Silberstein, R. A. (1981). Right Versus Privilege: The Open Admissions Experiment at the City University of New York. New York: Free Press.

Little, R. J. A., & Rubin, D. B. (1983). Missing data in large data sets. In T. Wright (Ed.), Statistical methods and the improvement of data quality. New York: Academic Press.

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

Little, R. J. A., & Schluchter, M. D. (1985). Maximum likelihood estimation for m ixed continuous and categorical data with missing values. Biometrika, 72(3), 497-512.

Lord, F. M. (1955). Estimation of parameters from incomplete data. Journal of the American Statistical Association, 50, 870-876.

Marini, M. M., Olsen, A. R. & Rubin, D. B. (1980). Maximum likelihood estimation in panel studies with missing data. In H. L. Costner (Ed.), Sociological Methodology (2nd ed.). San Francisco: Jossey-Bass.

Olkin, I. & Tate, R. F. (1960). Multivariate correlation models with mixed discrete and continuous variables. 448-465.

Orchard, T., & Woodbury, M. A. (1972). A missing information principle: Theory and applicatiobs. Proceedings of the sixth Berkeley Symposium on Mathematical Statistics and Probability, 1, 697-715. Berkeley: University of California.

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

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. Journal of American Statistical Association, 72, 538-543.

Rubin, D. B. (1974). Characterizing the estimation of parameters in incomplete-data problems. Journal of the American Statistical Association, 69, 467-474.

Rubin, D. B., & Thayer, D. (1978). Relating tests given to different samples. Psychometrika, 43(1), 3-10.

Sedransk, J. (1985). The objectives and practice of imputation. U.S. Bureau of the Census 1985, Proceedings of the First Annual Research Conference, pp. 445-452.

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. Scandanavian Journal of Statistics, 1, 49-58.

Trawinski, I. M., & Bargmann, R. E. (1964). Maximum likelihood estimation with incomplete multivariate data. Annals of Mathematical Statistics, 35, 647-657.

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