DOCUMENT RESUME

ED 374 130 TM 021 441

AUTHOR Huberty, Carl J.; Julian, Marc W.
TITLE An Ad Hoc Analysis Strategy with Missing Data.
PUB DATE Apr 94
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Adults; Case Studies; *Discriminant Analysis; Drinking; Foreign Countries; National Surveys; *Observation; Prediction; *Research Methodology; Statistical Analysis
IDENTIFIERS *Ad Hoc Strategy; Canada; *Missing Data

ABSTRACT

A subset of a real data set from a national survey on alcohol use and driving in Canada (original sample of 6,457) is used to illustrate an ad hoc analysis with missing data on multiple response variables. A complete-case analysis initiates this strategy, determining variables that may be deleted without losing effects of interest. By such a deletion, the number of complete observation vectors may very well increase. Also illustrated are two straightforward imputed means analyses. All illustrations are given in the context of predictive discriminant analysis. As discussed, the ad hoc strategy may be applicable to other multivariate contexts. Four tables illustrate steps in the analysis. (Contains 11 references.) (SLD)

An Ad Hoc Analysis Strategy with Missing Data

Carl J Huberty Marc W. Julian

University of Georgia

Paper presented at the annual meeting of the American Educational Research Association, April 1994, New Orleans.

Abstract

A subset of a real data set is used to illustrate an ad hoc analysis with missing data on multiple response variables. This strategy is initiated with a complete-case analysis to determine some variable(s) that may be deleted with no loss in effects of interest. By such a deletion, the number of complete observation vectors may very well increase. Also illustrated are two straightforward imputed means analyses. All illustrations are given in the context of predictive discriminant analysis.

An Ad Hoc Analysis Strategy with Missing Data

The purpose of this paper is to present an ad hoc analysis strategy with missing data for use in a predictive discriminant analysis context, and to illustrate the strategy using a subset of a real data set. Following the illustration, the same subset is subjected to two imputed means analyses, again for illustrative purposes.

Before presenting and illustrating the proposed analysis strategy, the missing data problem is briefly reviewed. Consider a data matrix of N (total number of experimental units) rows and p+1 (number of response variables plus one grouping variable) columns. Sometimes with real data sets there are fewer than the total possible N·p response variable measures; this is an example of a missing data problem. If there is a row with more than, say, p/2 missing measures, then this row (i.e., sampling unit) might be deleted outright. The same goes for a column (i.e., variable) with more than N/2 missing measures. Even subsequent to such deletions there may be additional missing measures, and the researcher needs to utilize the "working" data matrix (of N rows) in some analysis. One possible analysis is an available-case analysis. With this analysis, parameter estimates are determined using all available data; this involves using partially complete data matrix rows and columns. Such an analysis is very complicated and is outside the domain of most readily available computer software (see Hand, 1981, pp. 191-193).

Two analysis strategies often considered by practicing researchers with the final data set of interest are: (1) conduct a complete-case analysis; and (2) impute the missing data values and then conduct the analysis of choice. The complete-case analysis involves only those rows--say N* (< N) rows--in which there are no missing data. This strategy may be appropriate when the percent of rows with missing values is "low." It is the default analysis for most data analysis computer software (e.g., SAS and SPSS). [The percent of missing values may be calculated for the total data matrix (N·p variable measures), or for each group of units.]

Data imputation can be fairly simple or fairly complex. Four methods of imputation in a predictive discriminant analysis (PDA) context that have been studied are:

(1) Use the mean of all available scores in the respective groups or use the total-group means (see Hufnagel, 1988; Jackson, 1968);

(2) Standardize the scores (with zero means) for each variable; substitute zeros for missing values and conduct a principal component analysis of the pooled covariance matrix; coordinates of the point nearest the first principal component are substituted for missing coordinates (Chan, Gilman, & Dunn, 1976);

(3) Use the expectation-maximization (EM) algorithm, as described by Johnson and Wichern (1992, pp. 202-207) and Twedt and Gill (1992); and

(4) Let the missing value of the ith predictor, X_i , be determined as in (1) and similarly replace missing values for all other X's; then using X_i as a criterion variable, conduct a multiple regression with the other X's as predictors and use $\hat{X}_i$ as an imputed value; as described by Hufnagel (1988), extensive iterations may be incorporated.

Over the past 25 years or so, a number of data simulation studies have been reported that compare methods of handling missing data in a PDA context. Most of these studies focused on the two-group PDA situation. Four two-group studies spanning the last four decades are now briefly reviewed. Jackson (1968) compared the complete-case method and methods (1) and (4) mentioned above, and concluded that method (4) was best (in the sense of predictive accuracy), but not appreciably better than method (1) -- it was not clear as to which mean was used. Chan et al (1976) found that the complete-case method was "surprisingly good" (p. 844) for p = 2 and p = 4, and "it is comforting to find that [method (1) with separate-group means] performs reasonably well" (p. 844) relative to methods (2) and (4). The conclusion Hufnagel (1988) drew were conditioned on predictor intercorrelations, number(p) of predictors, and proportion of missing observations. It was concluded that, "in the case of large correlations [the complete-case method and method (4)] can be recommended best" but [method (2)]" could be used ... if small proportions of missing observations are given" (p. 74). Twedt and Gill (1992) concluded from their simulations that differences among methods (1) (which mean used is not clear), (2), and (3) were slight, and that it is "better to replace missing data than to delete the observation vectors with missing data" (p. 1577).

Group membership prediction rules typically used are normal-based rules. The effect of (random or nonrandom) missing data on multivariate normality is a complicated issue that is not discussed here (see Murty & Federer, 1991). In the ensuing discussion it is assumed that multivariate normality is a condition that is reasonably met.

An Ad Hoc Analysis

With the proposed ad hoc analysis, no data imputation is involved. This strategy may be described in the context of predictive discriminant analysis (PDA) as follows. Let there be N* (< N) units for which there are <u>complete</u> p-dimensional observation vectors. With an N^*-by-(p+1) data matrix, a complete-case PDA is conducted. From this analysis, it may be reasonable to conclude that one of the p predictor variables may be deleted with little loss, or in fact a gain, in predictive accuracy. Moreover, there may be units with missing data on the deleted predictor, but with complete data on the other p-1 predictors. One can then return to the original data matrix and determine a new data matrix of N** rows and (p+1)-1 columns, where $N^* < N^{**} \leq N$. Another PDA could then be conducted using the p-1 predictors and the N** units, and determine if another predictor might be deleted with little loss, or a gain, in predictive accuracy. If so, a new data matrix with, presumably, a greater number of rows than N** may then be analyzed. Again, a "weak" variable may be deleted; and so on. It is

recognized that multiple decisions would need to be made on different data matrices. Needless to say, judgment and reasonableness would need to be exercised. [Sometimes it may be judged reasonable to delete more than one predictor at a time. Suppose one considers deleting two predictors. Then what might be done is to conduct $\begin{bmatrix} p \\ 2 \end{bmatrix}$ analyses to determine which pair of predictors could be advisably deleted. Continuing to look for the next two predictors to drop, one would conduct $\begin{bmatrix} p-2 \\ 2 \end{bmatrix}$ analyses; and so on.]

An Example

Data. A data set was obtained via a national telephone survey in Canada. The survey dealt with alcohol consumption and automobile driving. Two types of drivers were determined: Group 1 consisted of those who did drive after drinking; and Group 2 consisted of those who did not drive after drinking. The purpose of one study that utilized the data set (DeJoy, Huberty, & Shewokis, 1993) was to develop a rule to predict group membership and to assess the predictive accuracy of the rule, particularly for Group 1. Thirteen predictor variables were considered. The initial data matrix had 6816 rows and 14 columns; one column was simply a group membership indicator. Four 13x13 correlation matrices were determined, two (list-wise and pair-wise deletion of units) for each group. Because no bivariate correlations were judged to be so high as to conclude there was extensive variable redundancy, all 13 predictors were retained for study. A total of 1705 drivers comprised Group 1 whereas 5111 drivers were in Group 2.

There were 359 drivers (62 in Group 1 and 297 in Group 2) who had missing score values on at least one of the 13 predictors. For the 359 drivers, 80 scores (9.9%) were missing in Group 1 and 568 scores (14.7%) were missing in Group 2. There were five variables for which there were no missing values. The variable with the largest number of missing values had 34 missing values in Group 1 and 172 in Group 2.

Prior to selecting the additional drivers from each original group, potential outliers were identified by using the SPSS DISCRIMINANT program (Release 4.0). A quadratic classification rule was used with respective group priors of .30 and .70 (see the next subsection). If a driver assigned to a group isn't very likely to have originated from that group as determined by a typicality probability of less than .05, then that driver was considered a potential outlier. Such drivers (65 in Group 1 and 6 in Group 2) were excluded from consideration in the selection process. To complete the data subset used for the current report, 100 drivers were randomly selected from the 1578 (nonoutliers) in Group 1 with complete observation vectors, and 200 were randomly selected from the 4808 (nonoutliers) in Group 2. This gives a final count of 162 Group 1 drivers and 497 Group 2 drivers. [These counts easily satisfy a rule-of-thumb sample-size guide of having at least 5 times the number of predictors in the smaller group, and the counts retain the approximate proportional sampling for the original 6816 drivers.] With these final group sizes, about 3.8% of the 162(13) Group 1 values are missing and about 8.8% of the 497(13) Group 2 values are missing; across the two groups of 659(13) values, about 7.6% are missing. Counts for the original sample and for the subset selected for this study are summarized in Table 1.

Table 1
Group Counts

	Group 1	Group 2	
Original Units			
Complete vectors	1643	4814	6457
(Outliers	65	6)	
Incomplete vectors	62	297	359
	1705	5111	6816
Selected Units			
Complete vectors	100	200	300
Incomplete vectors	62	297	359
	162	497	659

Analysis. It was necessary to make some decisions about the specifics of the PDA techniques to be used. For the group sizes involved and the two group covariance matrices (with outliers deleted the Box test yielded $P \doteq .0000$), it was decided that a quadratic classification rule was to be used (see Huberty, in press, Section IV-3). On the basis of familiar previous research data, prior probabilities of .30 for Group 1 and .70 for Group 2 were judged to be reasonable. In estimating proportions of correct classification (i.e., hit rates), in-doubt drivers -- those who are not clearly predicted to be one type of driver or the other -- were to be deleted from consideration. A threshold posterior probability of .60 was utilized; a driver had to "yield" a posterior probability of group membership of at least .60 to be assigned to one group or the other. Finally, an external analysis -- crossvalidation or leave-one-out (L-O-O) -- was used in estimating group hit rates. In sum, what was used is an externally assessed two-group quadratic classification rule with respective priors of .30 and .70 and a threshold value of .60. The hit rate of interest for the current study is that for Group 1. [See Huberty (in press, Section VI-3) for details of hit rate estimation.] The DISCRIM procedure in the SAS package (Release 6.07) was used.

A quadratic L-O-O rule was employed to order the predictors with respect to their relative contribution to classification accuracy for each group. To start, 13 12-variable analyses were run and the separate-group hit rates were noted with each predictor deleted. [Each of the 13 analyses was conducted using the drivers for whom there were complete observation vectors. Thus, the number of drivers considered may vary across the 13 analyses.] If a large hit rate for Group 1 (relative to the 13-variable hit rate) is associated with a deleted predictor, then that predictor is judged to be relatively unimportant. Then, after one predictor is deleted, 12 11-variable analyses were completed; and so on. Judgment must be exercised when hit rates associated with two or more deleted variables are "close."

Results. A summary of all steps in the analysis process is given in Table 2. The 13 12-variable analyses indicated that deleting V10 would actually increase the Group 1 L-O-O hit rate (from .460 to .512). As is obvious, by deleting the least important variable, V10, the number of rows in the complete-case matrix increases from 300 to 420. It turns out that by deleting three of the 13 predictors (one at a time), the number of drivers for whom complete score vectors were available increased from 300 to 461 while the Group 1 hit rate "stabilized" at about .53. By deleting a fourth variable, no appreciable increase in the number of complete vectors resulted without a drop in the Group 1 hit rate. So, with this data set, one could reasonably utilize a complete-case data set having 461 rows out of a total of 659 rows with the ad hoc analysis strategy. In utilizing this data set there would be a 28% increase in the number of Group 1 drivers, a 66.5% increase in the number of Group 2 drivers, and a 53.7% increase in the total number of drivers.

One might ask: "Why not use an analysis that utilizes all of the drivers on whom you have partial or complete observation vectors?" To do so, one could use some method of data imputation. We discuss such an approach next.

Imputed Means Analyses

As mentioned earlier, two types of means may be used for imputation purposes: total-group means and separate-group means. We judge the latter to better approximate the "real" observations (that are missing) and, therefore, focus on them. As indicated in the brief review in the introduction of this paper, the simple imputation method of replacing missing observations with means fares pretty well when compared with more complicated imputation methods (at least for the two-group context).

Imputed means may be utilized in a predictive discriminant analysis in two ways: (i) Impute all missing observations using means to complete the data matrix, determine a rule using this completed data matrix, and assess the rule using a L-O-O analysis; and (ii) Determine a classification rule using only the complete observation vectors, and then apply the rule to the data matrix with all missing observations imputed with means so as to arrive at hit rate estimates.

Both imputed means analyses were conducted on the data set described above, using the SAS package (Release to 6.07).

Imputed Means Analysis (i). Separate-group means were calculated using available data for each response variable for which there were some missing scores. This was accomplished using the MEANS procedure from SAS. In the original N*(p+1) data matrix, each missing data point for a predictor was supplanted by the corresponding group mean of the predictor. The resulting augmented matrix data matrix (N=659) was used for the following analyses. First, a 13-predictor PDA was conducted (N = 659) using a quadratic L-O-O analysis. Second, 13 12-predictor analyses were conducted to determine if, by deleting a predictor,

Table 2
Results of Ad Hoc Classification Analyses

No. Predictors	Predictor Deleted	Group 1 L-0-0 Hit Rate	Complete Case Nos. G_1	G_2	Total
13	(none)	.460	100	200	300
12	V10	.512	125	295	420
11	V4	.526	127	327	454
10	V7	.531	128	333	461

an increase in the Group 1 hit rate (relative to the 13-predictor hit rate) would result. It turned out that by deleting V7 the Group 1 L-0-0 hit rate increased from .506 to .518. Third, 12 11-predictor analyses were conducted. It turned out that by deleting V2 or V4 or V9, the Group 1 L-0-0 hit rate remained at .518. Thus, three sets of 11 10-predictor analyses were conducted, one set with V7 and V2 deleted, one with V7 and V4 deleted, and one with V7 and V9 deleted. It turned out that by deleting that last pair along with V1, the Group 1 L-0-0 hit rate increased to .525 -- this was a greater increase than those resulting from the sets of analyses with the other two pairs deleted. A summary of the analyses is presented in Table 3.

Imputed Means Analysis (ii). With this method, a classification rule is built using the 300 complete observation vectors. Then this rule is applied to the data matrix with 659 rows wherein separate-group means replace the respective missing variable values. This type of analysis may be carried out using the SPSS DISCRIMINANT

Table 3

Results of Imputed Means Analysis (i)

No. Predictors	Predictor Deleted	Group 1 L-0-0 Hit Rate
13	(none)	.506
12	V7	.518
11	V9	.518
10	V1	.525

program (with the keyword MEANSUBSTITUTION), except that total-group instead of separate-group means would be used as imputed values. Furthermore, SPSS is unfortunately limited to a linear internal or resubstitution classification analysis which would be inappropriate for this particular set of data. Using the "TESTDATA" option in SAS DISCRIM it is straightforward to develop a classification rule based on complete observations and subsequently apply the rule to the data matrix augmented by separate-group means. Notwithstanding, it is not possible to obtain L-O-O results via SAS DISCRIM in this context. Despite this shortcoming, the quadratic internal analysis approach was considered as an alternative. Of course, the obtained hit rates should not be compared with those obtained via the two previous analysis approaches.

Results of this imputed means analysis of the current data set are given in Table 4. It is not too surprising that the internal hit rates obtained for 13, 12, and 11 predictors were higher than the corresponding L-O-O hit rates using analysis (i) and using the ad hoc analysis.

Discussion

The intent of presenting the three analysis strategies was not to compare them in any empirical sense. These are simply three strategies that are fairly easy to carry out with practically any data set in the illustrated context of a predictive discriminant analysis. Although the strategies were illustrated in a two-group situation, they may be applied to a situation involving three or more groups in a similar manner.

The ad hoc strategy in general may be applicable in other multivariate contexts; for example, multivariate analysis of variance (MANOVA), multiple regression analysis, and descriptive discriminant analysis. If such a missing data analysis strategy were used in a MANOVA context, two questions would need to be addressed: (1) What interpretation of outcome variable relative importance is to be considered?; and (2) What numerical index of relative importance, consonant with the selected interpretation, is to be used? Answers

Table 4

Results of Imputed Means Analysis (ii)

No. Predictors	Predictor Deleted	Group 1 Hit Rate
13	(none)	.531
12	V4	.543
11	V9	.549

to these questions would presumably determine the dimensions of the data sets to be used in the analysis strategy. With the data set used earlier in this paper, for example, it would have to be described if the initial analyses should be based on a data matrix with 300 rows or with varying number of rows if multiple analyses are carried out in the process. The above two questions would also have to be considered in other analysis contexts -- see Huberty (1989) for a discussion on variable ordering.

The general philosophy behind the proposed ad hoc analysis strategy is: Do the best with what you have. Some might argue that the available data may be utilized in estimating missing scores which would result in a more acceptable analysis. Perhaps. Some data imputation methods assume randomly missing data. In a given study, how "good" imputed data are is virtually unknown. Another type of question that may be asked in a context like that used herein is: What is to say that a variable would be assessed as an unimportant predictor if more measures on the variable were available? Of course, an ad hoc analysis strategy such as that proposed here is expected to work with varying proficiency across different data sets. It may very well be an analysis strategy of choice for some real data sets.

References

Chan, L. S., Gilman, J. A., & Dunn, O. J. (1976). Alternative approaches to missing values in discriminant analysis. Journal of the American Statistical Association, 71, 842-844.

DeJoy, D. M., Huberty, C. J, & Shewokis, P. A. (1993, May). Attitudinal and behavioral predictors of drinking and driving decisions: Preliminary analyses of a large-scale survey. Paper presented at the second World Injury Control Conference, Atlanta.

Hand, D. J. (1981). Discrimination and classification. New York: Wiley.

Huberty, C. J (1989). Problems with stepwise methods: Better alternatives. In B. Thompson (Ed.), Advances in social science methodology (vol. 1, pp. 43-70). Greenwich, CT: JAI Press.

Huberty, C. J (in press). Applied discriminant analysis. New York: Wiley.

Hufnagel, G. (1988). On estimating missing values in linear discriminant analysis - Part I. Biometrical Journal, 30, 69-75.

Jackson, E. C. (1968). Missing values in linear multiple discriminant analysis. Biometrics, 24, 835-844.

Johnson, R. A., & Wichern, D. W. (1992). Applied multivariate statistical analysis. Englewood Cliffs, NJ: Prentice Hall.

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

Murty, B. R., & Federer, W. T. (1991). Missing observations in multivariate analysis. Journal of the Indian Society of Agricultural Statistics, 43, 107-126.

Twedt, D. J., & Gill, D. S. (1992). Comparison of algorithms for replacing missing data in discriminant analysis. Communications in Statistics -- Theory and Methods, 21, 1567-1578.