ANALYZING SURVEY DATA BY COMPLETE-CASE AND AVAILABLE-CASE METHODS

Michael P. Cohen, Bureau of Transportation Statistics and Gary G. Huang, Synectics for Management Decisions
Michael P. Cohen, 400 Seventh Street SW #3430, Washington DC 20590

Key Words: Casewise deletion, Listwise deletion, Longitudinal surveys, Item nonresponse, Unknown category

Abstract: On many data files, cases are left missing rather than imputed when there is item nonresponse. This is particularly common for longitudinal surveys. Common ways handling the item nonresponse when doing analyses based on such files are to use complete cases only (listwise deletion) or available cases (casewise or pairwise deletion). The advantages and pitfalls of these methods are explored through examples and discussion.

1. Introduction

Item nonresponse (hereafter referred to as missing data) is virtually inevitable in survey data collections. It poses problems for survey data analysis, especially in multivariate statistical analyses such as least square regression, logistic regression, factor analysis, event history analysis, and hierarchical linear modeling. This is because missing-data structures are often multivariate in the sense that the missing values are not confined to a single variable. With incomplete data, it is difficult to estimate the sample mean vector and sample covariance matrix, the basis for initial data reduction in multivariate analysis (Little and Rubin, 1987). Further, the number of missing cases increases quickly as multiple independent variables are entered into equations.

With increasing sophistication, imputation has become a valuable approach to dealing with the problem (e.g., Kalton and Kasprzyk, 1982; Little, 1986; Little and Rubin, 1987; Schafer, 1997; Olsen and Schafer, 1998). However, statistical agencies do not generally conduct imputation for every survey dataset they have produced. With technical complications and costs, imputation is not generally feasible for secondary data analysts (Wang, Sedransk, and Jinn, 1992) who generate a large bulk of research in both private and government sectors. In practice, various ad hoc approaches to reduce the problems in multivariate analyses are common in the literature, including government statistical reports. But such procedures are often inconsistent and nonsystematic and may generate confusing and misleading research results. In this article, we study this issue and make recommendations.

2. Literature Review

To learn about how the problem is theoretically and practically handled, one needs to examine systematically the current literature of survey-based research and subject matter research (in this article, we concentrate on education research). Our focus was the treatment of *item nonresponse* in multivariate analytical procedures. Issues of missing data resulting from *unit nonresponse*, including various forms of sample deficiency, sample attrition, and data censoring, are not in the scope of the search. Specifically, we reviewed statistical and methodological journal articles and monographs in an attempt to overview the theoretical treatment of the topic. We also reviewed and evaluated empirical education research that applied multivariate statistical procedures to analyze survey data. This literature allows us to see how current research in practice handled the missing data without imputation.

2.1 Applied Statistical and Methodological Literature

We searched the Current Index of Statistics (CIS) for publications on item nonresponse and missing data in survey analysis. The formal expression and assessment of the effect of nonresponse concerned early researchers on missing data-related issues. First, it was necessary to estimate the bias in commonly used statistics (e.g., the sample mean of a variable) resulting from data with item nonresponse. This could be done by using Bayesian techniques to calculate a probability interval for the statistics such that the "subjective notion" of nonresponse effect can be formalized (Rubin, 1977). The probability interval is conditional on the observed data. To refine the test of missing completely at random (MCAR) with multivariate missing data, researchers proposed a single global test statistic for MCAR (Little, 1988). This test, using all of the available data, provided the asymptotic null distribution and the small-sample null distribution for multivariate normal data with a monotone pattern of missing data. The test reduced to a standard t test when the data were bivariate with missing data limited to a single variable, and the results seemed conservative for small samples (Little, 1988).

Evaluative comparisons of different methods frequently used by secondary analysts help identify the relative weakness and strength of the methods in applied settings. The resulting recommendations could improve secondary data analyses. Focusing on the interplay of different imputation techniques and different methods used by secondary data analysts, Wang, Sedransk, and Jinn (1992) examined a number of options, including using only the observed values, mean imputation overall, random imputation overall, simple regression imputation, random regression imputation, and random imputation within adjustment cells. Under the assumption of missing at random (MAR), the researchers used simulated data to examine the confidence intervals for regression coefficients in linear regression analysis. The results indicated that general-purpose imputation methods such as mean imputation overall and random imputation overall were not acceptable. Multiple imputation that incorporated information about the nonresponse process was recommended (Wang, Sedransk, and Jinn, 1992).

Further efforts were made to explore the complication of analysis of data with missing values collected in a complex survey. With a stratified or clustered sample design, such surveys generate data that require design adjustment in estimation, a procedure that modifies the standard statistical estimation. With non-ignorable missing data, the assumption underlying the design adjustment becomes problematic. Some studies suggested censoring models to account for the nonresponse process and to incorporate information about the design effect in design adjustment and estimation (Chambers, 1988).

Research has generated recommended practices dealing with different aspects of missing data-related problems. For example, proposed techniques cover issues of regression estimation for categorical variables, simultaneous use of response model and parametric model to protect bias in estimation, and selecting auxiliary variables to improve precision of estimates (Kott, 1994; Skinner and Nascimento Silva, 1997). Although the statistics literature apparently stresses imputation and estimation techniques, selected ones did provide secondary analysts with some insights and options that may work in appropriate analytical settings.

We also searched the Education Resource and Information Center (ERIC) database—the world's largest educational bibliographic database—for statistical and methodological research publications addressing multiple regression analysis of survey data involving missing values, published since 1995. In addition, we searched the online catalogs at the University of Maryland (College Park) and the Georgetown University libraries for monographs on the same topic published since 1995.

Our ERIC search generated a listing of journal articles and conference papers. The dominant approach used in the literature was imputation methods. There were only a few studies concerning the impact of missing values in survey or test data analysis and the varying consequences of deleting missing values. Ad hoc analysis of survey data with missing values was examined by a small number of reports.

2.2 Empirical Education Research with Multivariate Analysis

To scrutinize the issue of missing data in multiple regression and related procedures, we reviewed most recent National Center for Education Statistics (NCES) reports. We also reviewed 1999-2000 publications of survey-based multiple regression analyses in two major education research journals: *Sociology of Education* and *Educational Evaluation and Policy Analysis*. We gathered information about the ways missing data were described and treated in the analyses.

NCES reports do not regularly discuss the missing data issue. Only two reviewed reports mentioned "pairwise deletion of missing data" with the Data Analysis System (DAS) that generated correlation matrices for further calculation of regression coefficient estimates.

Journal articles are more likely than NCES reports to present missing-data-related information in multivariate analyses. Our review suggests that data analysts did not deal with the missing data consistently—or at least did not show so in their research presentations. Their approaches can be roughly grouped in three large categories in terms of clear description of missingness (scope, pattern, and mechanism); reasonable treatment (missing case deletion, missing case flagging, examining the impact of missing data, and imputation); and systematic presentation of material in the report. The "Comprehensive Group" includes studies that handled well the three aspects. The "Attentive Group" covers studies that were conscious of the issues and involved efforts to cope with the problems but did reach a satisfactory level. The "Neglectful Group" refers to studies that did not make substantial effort to work on any of the aspects. Substantial numbers of articles fell into each of these three groups.

3. Suggested Approaches

In applied research using multivariate procedures to analyze survey data, analysts need to cope with missing data problems in ways that are both systematic and feasible. The term *systematic* refers to the thoroughness and logical rigor in examining the missing data, testing and using appropriate strategies, and presenting the results. *Feasibility* refers to the likelihood of successful completion of such tasks by secondary data analysts, perhaps in terms of the complexity of the process and the amount of effort.

By assessing different approaches with the two concerns in mind, we propose some balanced recommendations for developing workable and reasonable strategies to reduce the bias resulting from missing data. We provide general rationale by synthesizing information largely from relevant chapters in two monographs on applied multivariate analyses with missing data (e.g., Chapter 7 in Cohen and Cohen, 1983; Chapter 3 in Little and Rubin, 1987).

3.1 General Rationale

To understand why missing data cannot be used in multivariate analyses without appropriate treatment, consider a rectangular (n by K) data matrix $\mathbf{X} = (x_{ij})$, where x_{ij} is the value of a variable X_j for observation i, $i = 1,\ldots,n$, $j = 1,\ldots,K$. Without missing data, multivariate analyses typically entail initial reduction of data to the vector of sample means

$$\bar{x} = (\bar{x}_1,\ldots,\bar{x}_K)$$

and the sample covariance matrix $\mathbf{S} = (s_{jk})$, where

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k).$$

When missing data exist, the estimation of $\bar{x}$ and $\mathbf{S}$ becomes problematic (Little and Rubin, 1987).

A critical question facing analysts is to determine whether the missing data are at random (Little and Rubin, 1987; Cohen and Cohen, 1983). When data are missing at random, the presence or absence of data is not related to the variables under study, at least not related to the missing data. For example, given $K = 2$, X_1 (age) and X_2 (income) are two variables in the analysis. Either X_1 or X_2 may be missing and missingness is not related to either X_1 or X_2. But this is typically not the case in practice. Usually in survey data, high- and low-income respondents are more likely to have missing values than middle-income respondents, i.e., missingness is related to X_2 (the missing data) though not to X_1. If excluding missing data from the analysis, the marginal distributions of age and income with only the complete cases are distorted by the overrepresentation of middle-income respondents. Estimation of the correlation between X_1 and X_2 may be biased. Estimation of parameters of the linear regression of X_2 on X_1 from the remaining complete cases may be also biased, though not so with the regression of X_1 on X_2.

Analysts frequently encounter missing data that are related to variables under study. For example, nonresponse to items on student motivation is related to student motivation itself: poorly motivated students are less likely than other students to respond to the question. The missingness on motivation could also be related to school experience, behavioral and emotional problems, socioeconomic status, race-ethnicity, and other variables of interest in research. To make decisions in treating missing data, analysts must answer the question of whether the factors that cause nonresponse overlap with factors that are under study—both dependent and independent variables—in the multivariate analyses (Cohen and Cohen, 1983).

This leads to another concern: If the *dependent* variable—X_2 in the above analysis—has missing values, the risk of losing information is high after excluding the missing cases. Furthermore, deleting cases with missing data on X_2 may cause the remaining sample no longer to be representative of the population that the original sample was designed to represent (Cohen and Cohen, 1983).

A related practical issue is the extent of missing data. Two measures are important here: The proportion of missing data on either X_1 or X_2, p_a, and the sample size n in the analysis. Generally, if p_a is small and n is large, then the problem of missing data is obviously less damaging—even if the missingness is not at random (Cohen and Cohen, 1983). If p_a is large or n is small, one must examine data carefully to learn why the data are missing and whether missingness is at random.

3.2 A Step-by-Step Approach

Other than model-based imputation, analysts in practice frequently use three forms of data deletion to cope with missing data in multivariate analyses, namely, deleting the missing cases pairwise, deleting missing cases listwise, and deleting the variables with missing values. The advisability of data deletion, however, is controversial. Data deletion generally runs into risk of wasting information, reducing statistical analytical power, lowering precision of estimates, and excessively relying on the assumption of the complete randomness of missingness, which, if unwarranted, generates biased or nonrepresentative statistics (Cohen and Cohen, 1983).

There are more recommendable ways to deal with missing data, requiring relatively complicated processes. With different circumstances under which substantive research is conducted, there probably is no single best way to handle the problems. Analysts may consider using alternative approaches and data deletion to cope with the missing data-caused problems. Therefore, rather than presenting discrete methods, we suggest some steps for approaching missing data issues based on balanced consideration of potentially workable methods. Intended to be both systematical and manageable, we propose these steps specifically for secondary data analysts who cannot conduct model-based imputation of missing data because of the technical complication and high costs.

3.3 Examination of Data Missingness

Several issues must be examined regarding missing data, namely, randomly vs. selectively missing data, the extent of missing data, and the pattern of missing data.

Randomly vs. selectively missing data: First, analysts must try to understand why missing data happen and to determine whether data are missing randomly or selectively. For listwise deletion of missing cases (that is, complete case analysis), which is often used as a quick coping strategy in applied research, one can use a simple procedure to examine the missing-at-random assumption. It is to compare the distribution of a variable X_j based on cases after the listwise deletion (i.e., only retain cases that have no missing data on any variables) with the distribution of X_j based on all cases for which X_j is recorded. Significant differences in descriptive statistics (means and standard deviations) indicate the missing-at-random assumption is invalid and the listwise deletion—or sometimes called analysis with complete cases—generates biased estimates (Little and Rubin, 1987). This type of test should work fairly well with large national survey data with a large sample size. A test of missing-at-random for specific variables in multivariate analyses requires constructing and including a specially coded missing indicator for X_j into the equation. The resulting coefficient estimate and significance test allow analysts to either reject or accept the null hypothesis that missingness is at random.

Extent of missing data: The analyst should examine the proportion of missing data p_a on each variable under study, including independent variables and the dependent variable. The overall extent of the missing data also should be considered. With a substantial number of variables in the multivariate analysis, each is missing or observed independently according to a Bernoulli process with even a small chance of missingness, the expected proportion of complete cases can be surprisingly small. Although in reality missingness may to some extent overlap between variables, the more variables involved in the analysis, the more extensive the overall missing data. The extent of missingness should be always assessed in connection with the sample size.

Pattern of missing data: Next, the analyst may examine the pattern of missing data. In addition to assessing missing data by variables, one may sort out the concentration of missing data by respondents, i.e., certain respondents have missing data on a large number of variables, whereas the rest have complete data or have missing data on only a small number of variables. If the fraction of nonrespondents is small enough and there are reasons to argue that the cause of nonresponse is not related to the relationships being studied, then identifying missing data concentration by respondents may help make the decision to delete these cases.

3.4 To Delete or Not to Delete?

Although standard software packages such as SAS and SPSS feature either listwise or pairwise deletion of missing cases, using deletion is questionable if the missing-at-random assumption is not warranted. For example, in pairewise deletion, a matrix is computed for each pair of variables (X_i, X_j) by using cases with values on both variables. If data are missing on either X_i or X_j or both for a case, this case is deleted from the computation of the correlation coefficient r_{ij}. Thus, the resulting r_{ij} is not based on a sample different from the original one. If for each pair of variables a different number of missing cases are deleted, then the resulting r_{ij} are based on different samples. The correlation matrix, together with sample means and the standard deviation matrix (with mean and standard deviation for available cases for each X_i) is then used to estimate other statistics of the regression equation (Cohen and Cohen, 1983). This method of deletion is also called the available-case method (Little and Rubin, 1987). Note that these statistics are representative of the targeted population only if the data are missing at random. Even so, the results make the statistical inference awkward because varying sample sizes are involved in estimation.

Listwise deletion runs in a slightly different way. The difference is that when data are missing on *any* variable for a case, this case is then deleted from the computation of the correlation coefficients of not only the given pairs of variables (X_i, X_j), but also of all the other pairs of variables under study. The consequence is equivalent to analysis with complete cases, i.e., estimation is based on a same subsample of complete cases from the original sample (Little and Rubin, 1987). An obvious advantage of listwise deletion is that the resulting estimates are not a hodgepodge with numerous different subsets of data with different sample sizes. With the same subsample used in estimation with listwise deletion, however, the analyst still faces the danger of estimation bias if the missing-at-random assumption is questionable. Also, because typically a large number of cases are deleted in the process, listwise deletion wastes more information than pairwise deletion.

A grave concern of bias arises if the missing-at-random assumption is false, which is often the case in survey data analysis. If missingness is related to the variables under study, the estimates from multivariate procedures refer to systematically different subgroups of the population and thus are hardly interpretable.

Analysts may consider dropping variables that contain a large number of cases with missing data. This would not be a bad option if the variable can be reasonably seen as contributing little to accounting for the variance of the dependent variable. In fact, excluding such variables from the equation could increase the statistical significance and the precision of the regression coefficients and coefficient of determination. But, in most cases of practical analysis, variables under study must be thought to be conceptually important in the model, and removing them from the analysis due to missing data is always a loss of information and damaging to the conceptual framework that is critical to applied research.

In general, data deletion of some form by itself is not satisfactory. The problem inherent in data deletion is the failure to see missing data per se as potentially useful information for research. Ironically, the thorny issue of nonrandom (or selective) missingness sometimes can be more helpful in research than random missingness. Missing data on certain variables by certain respondents are facts that researchers should exploit in connection with their research questions. Instead of simply avoiding the problem as data deletion is intended to do, it is desirable and possible to represent the *absence* of data as predictors together with other variables to account for the criterion variance in multivariate analyses (Cohen and Cohen, 1983).

4. Conditional Missing Data

With the conditional skip pattern designed in a questionnaire survey, some respondents are expected to have missing data. For instance, in a survey of high school students, an item asks about whether respondents participated in advanced academic programs; those who responded "no" are expected to skip a following set of items about these programs because these items are not applicable to them. The resulting missing data for these respondents are called *conditional* missing data. A conventional way to analyze such data is to use only those cases that participated in the programs. With a number of such skip patterns in a survey, this subset approach may make the analysis fragmented and the interpretation cumbersome.

The item on program participation serves as a ready missing data indicator. For the non-participants, we could substitute for the missing data on each of the program variables the mean or a constant. Then data for the whole sample can be used in a single analysis, with the program participation item and any other independent variables of interest included.

5. When Not to Use the Missing-Data Indicator

Under certain circumstances, using the missing data indicator is unnecessary or even undesirable. Analysts should always carefully check the data to determine if it is appropriate to use the procedure.

The proportion of missing data

Generally, it is problematic to use the missing-data indicator, when the proportion of missing data in the sample, p_a, is very small or very large. This is especially true if the sample size, n, is also small. Clearly, when both p_a and n are small, the number of missing cases, n_a, will be small and the resulting estimate for the missing-data group's criterion mean, $\overline{Y}_a$, will be unreliable. Unless it is known that $\overline{Y}_a$ is far different from $\overline{Y}$, it is probably more sensible not to use the missing indicator. Oppositely, if p_a is very large, the number of the data-present cases will be small and the resulting estimates will be unreliable. Thus in either situation, analysts should consider ways other than using a missing indicator to cope with the missing data problem.

Multiple independent variables with missing data

Survey data often contain many variables with nonrandom missing data. If there are many such variables in the analysis, it would not be advisable to use a missing-data indicator for each of them. The multiple missing indicators are likely to be correlated substantially and carry little unique variance in relation to the dependent variable. The collective inclusion of these indicators in the equation will lead to unstable regression estimates due to the reduced error degrees of freedom and high correlation among independent variables. An alternative to deal with this problem is to create a single indicator to represent the respondents' general tendency to have missing data. Such an indicator can be scored with the number of items on which data are missing. It also can be scored with factor scores generated from a factor analysis of the matrix of binary missing data indicators. Of course, this can be done only if the factor analysis yields a single dominant factor. The procedure of a single combined missing indicator should be better than the procedures that do not include any missing indicator or that include multiple ones.

Missing-at-random assumption is valid

If the assumption that data are missing at random is known or can be confidently established, then either pairwise or listwise deletion of missing cases can be considered. It is thus unnecessary to use the missing indicator approach. Moreover, inclusion of missing indicators may even reduce statistical power and stability of the multivariate analyses (Cohen and Cohen, 1983). Caution should be always exercised, however, in making the randomness assumption.

6. Final Comments

With the proposed approach, the analyst essentially takes a pragmatic perspective regarding the missing data problems. A specially coded missing data indicator represents the absence or presence of data on a given variable. The absence of data is seen, not simply as a data flaw to be avoided, but as a fact that has potential value for investigation in relation to the dependent variable.

Acknowledgments

Michael P. Cohen worked for the National Center for Education Statistics (NCES) when most of this work was performed, and Gary G. Huang received financial assistance from NCES. Both authors are grateful to NCES and, in particular, to the Chief Statistician, Marilyn M. McMillen. This paper is intended to promote the exchange of ideas among statisticians and other researchers.

The views are those of the authors, and no official support by the U.S. Department of Education is intended or should be inferred.

References

Chambers, R. L. (1988),"Design-Adjusted Regression with Selectivity Bias." *Applied Statistics*, 37, 323-334.

Cohen, Jacob, and Cohen, Patricia. (1983), *Applied Multiple Regression Analysis/Correlation Analysis for the Behavioral Sciences*, (2nd Ed.), Mahwah, New Jersey: Lawrence Erlbaum.

Kalton, Graham, and Kasprzyk, Daniel (1986), "The Treatment of Missing Survey Data," *Survey Methodology*, 12, 1-16.

Kott, Phillip S. (1994), "A Note on Handling Nonresponse in Sample Surveys," *Journal of the American Statistical Association*, 89, 693-696.

Little, Roderick J. A. (1986), "Survey Nonresponse Adjustments for Estimates of Means," *International Statistical Review*, 54, 139-157.

Little, Roderick J. A. (1988), "A Test of Missing Completely at Random for Multivariate Data with Missing Values," *Journal of the American Statistical Association*, 83, 1198-1202.

Little, Roderick J. A., and Rubin, Donald B. (1987), *Statistical Analysis with Missing Data*, New York: Wiley.

Olsen, Maren K., and Schafer, Joseph L. (1998), "A Class of Models for Semicontinuous Longitudinal Data," American Statistical Association *Proceedings on Survey Research Methods*, 721-726.

Rubin, Donald B. (1977), "Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association*, 72, 538-543.

Schafer, Joseph L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman and Hall.

Skinner, C., and Nascimento Silva, P. D. L. (1997), "Variable Selection for Regression Estimation in the Presence of Nonresponse," American Statistical Association *Proceedings on Survey Research Methods*, 76-82.

Wang, R., Sedransk, J., and Jinn, J. H. (1992), "Secondary Data Analysis When There Are Missing Observations," *Journal of the American Statistical Association*, 87, 952-961.