ASSESSING THE EFFECTS OF MISSING DATA

John D. Hutcheson, Jr. and James E. Prather, Georgia State University

Problems resulting from incomplete data occur in almost every type of research, but survey research is especially prone to produce data sets within which some values for some subjects are missing. Many different methods for handling missing data have been proposed and employed. However, most commonly used procedures treat the missing data problem as "a disaster to be mitigated" rather than as a "pragmatic fact that may be investigated" (Cohen and Cohen, 1975: 288). In this paper, we discuss some of the problems associated with commonly used methods of handling missing data and illustrate the use of an alternative method, proposed by Jacob Cohen (1968: 438), that treats the fact that there are missing observations in a manner which permits the researcher to identify how missing observations could affect analysis and interpretation. Missing data are viewed, from the perspective explained below, as a specification problem rather than as a technical inconvenience.

Commonly Used Methods of Handling Missing Data

Most commonly used methods of handling missing data assume that missing observations occur randomly. When employing such procedures in survey research, the implicit assumption is that the fact that some respondents refuse to answer or are unable to respond to some questions is not related to any of the other items included in the analysis of the data. A similar assumption is used when it is assumed that refusals to participate as a respondent in surveys occur randomly. Nonrespondents are assumed to be similar to respondents or to differ only in ways unrelated to the content of the survey instrument. Hesseldenz (1976) has treated this type of nonresponse as a specification problem and has shown how the nature of nonresponse bias may be examined to enhance analysis and interpretation. The perspective and approaches employed below are similar. However, the problem addressed here is the effect of missing data for specific items rather than the effect of missing data for entire cases (respondents). Since most surveys include attitudinal items and/or questions related to personal characteristics, the respondent, while permitting the interview, may refuse or be unable to respond to specific questions. The assumption that this occurs in a manner which is not related to other variables included in the analysis of the data merits examination.

While assuming that missing observations occur randomly (Hertel, 1976; Gleason and Staelin, 1975; and Press and Scott, 1976), most commonly used methods for handling missing data can result in undesirable effects, in addition to incorporating this uninvestigated assumption. Case-wise (also referred to as list-wise) and pair-wise deletion of missing data, often employed in analyses using software packages such as the Statistical Package for the Social Sciences (SPSS), redefine the original sample to include what could be an unrepresentative subsample of the population and ignore factors that could be important in interpretation.

Many researchers have attempted to "plug" missing observations by substituting sample means for missing data. This procedure, again, assumes that missing observations occur randomly. Additionally, using the mean substitution method reduces the total variance observed and results in conservative estimates of association. These procedures fail to incorporate the informational value of missing data in a manner which allows the researcher to determine if analyses are biased (misspecified), and may, therefore, result in misinterpretation.

It has been argued that the "best one can do is select a missing data routine which does not increase biases already in the available data" (Hertel, 1976: 460). We contend, however, that the methods suggested by Cohen (1968: 438), and described, illustrated and expanded below, permit the researcher to assess the effects of missing data in a specific analysis and consequently minimize the possibility of undetected bias and misinterpretation. This, we believe, is more desirable than merely insuring that the existing bias is not increased.

An Alternate Method

The method described here allows the researcher to estimate, first, if missing data are systematically associated with substantive variables in a given analysis. If systematic relationships do exist, that is, if missing observations are related to other variables in the analysis, the researcher may then assess the nature of such bias and incorporate this additional information into the interpretation of the analysis.

The method which Cohen (1968: 438) has suggested and which is described in more detail in Cohen and Cohen (1975: 265-290) involves the substitution of sample means for missing observations. In contrast to the mean substitution method, however, this procedure concurrently employs the creation of dummy variables for every variable in which means have been substituted. For example, if a sample mean for education is 11.8 years and 50 observations of the 250 in the sample are missing, the value 11.8 would be used to "plug" the missing observations and a dummy variable, "missing education," would be created assigning a value of "0" to each actual response to the item and a value of "1" to each missing observation.

If education is an independent variable in a given analysis, a regression model may be used to identify systematic relationships of the actual responses to the education item and the missing education index with the dependent variable. If the analysis results in a significant relationship between the missing education index and the dependent variable, in the presence of the actual education variable, missing observations in education have not occurred randomly and the researcher may assess the nature and consequences of the resulting bias.

If, on the other hand, education is a dependent variable, with a limited number of categories, the researcher could create a missing education category to employ along with the other education categories so that discriminant analysis could be used to identify systematic differences between those subjects who responded to the education item and those who did not. Such procedures utilize all available information, including the fact that some respondents provide valid information and some do

not. This allows the researcher to assess the nature of the bias that may be introduced by missing data.

An Example

Participation of eligible voters in the electoral process has been considered a critical feature of traditional democratic theory. Consequently, many scholars have sought to explain why some eligible citizens vote and others do not. Much of what we know about voter participation is based upon survey data.[1] Research employing survey methods, for example, has found that participation is associated with education, income and other measures of social status (Verba and Nie, 1972: 125-137). Ben-Sira (1977) has synthesized a large body of research in formulating an explanation for the positive relationship between social status and voting. In general, it is thought that higher status is associated with greater resources (education, income) and that "the greater one's resources (or personal power potential) the greater one's striving for realization of this potential through achievement of a higher level of power" (Ben-Sira, 1977: 1970). The higher status person, then, has higher levels of political interest (Berelson, Lazarsfeld and McPhee, 1954: 25), political efficacy (Campbell, Converse, Miller and Stokes, 1964: 253) and a greater awareness "of the impact of government on the individual" (Ben-Sira, 1977: 1970).

A survey of over 7,000 residents (a 1% sampling) of Atlanta and Fulton County (part of suburban Atlanta), Georgia, conducted in mid 1976, provides an opportunity to explore some of the hypotheses posited above. This analysis, then, provides the context within which the consequences of employing differing procedures for handling missing data are illustrated. The 1976 Atlanta/Fulton County survey provides a data set with which to demonstrate these consequences when commonly used missing data procedures are most justifiable -- when the sample size is unusually large. The impacts of commonly used missing data routines, in most instances, will decrease as the size of the sample increases. Thus, if negative consequences are apparent in the analysis of the Atlanta/Fulton County survey data, they are likely to be quite severe when these routines are used with smaller data sets.

In this example, the dependent variable is voting, having voted in the past 5 years (1) or not (0). In the first analysis presented here, regression is used to identify bias attributable to missing observations in the independent variables. Explanation of voter participation is sought in terms of socio-economic variables (race, family income, education and age) and political attitudes (political efficacy and interest and governmental salience). In accord with the above theory, education, family income and political efficacy and interest are hypothesized to have positive effects on voter participation, as is whether or not the respondent felt that government had a "good deal" of impact on him/her, as an individual (governmental salience). Race and age are included in the model as control variables. Non-white was coded "0" and White was coded "1', since previous research suggests a positive association between "being White" and voter participation (Campbell, Converse, Miller and Stokes, 1964: 150). Since respondents in the survey were as young as 18 years of age, and because voting and non-voting were operationally defined to include the 5 years prior to the survey, the necessity to control for age is apparent.[2]

The voter participation model was estimated using four different specifications for missing data (see Table 1). The case-wise (or list-wise) deletion routine eliminates any case in which there are missing observations for any of the variables included in the model. The pair-wise deletion technique causes any case with a missing observation for a particular variable to be deleted from calculations involving that variable only. The mean substitution specification, as described earlier, "plugs" missing observations for variables with sample means. The fourth missing data specification is the mean substitution and dummy variable method described in the previous section.[3]

Table 1 provides the regression estimators and standard errors for each of the four specifications of the voter participation model. From an examination of the missing data variables in the fourth specification, it is clear that missing observations in the education, age, political efficacy and governmental salience variables are systematically related to voting. Only one of the three missing family income variables is significantly related to voter participation; while refusing to provide income information and missing income (no reason given) do not occur systematically, not being able to provide family income data ("don't know family income") is systematically related to voting.[4]

While the estimators and standard errors in the four specifications are fairly consistent, the lower R^2's for the case-wise (list-wise) deletion and the mean substitution specifications reflect the decrease in variance attributable to loss of cases and degrees of freedom in the case-wise (list-wise) deletion specification and to loss of variance in the mean substitution specification. Thus, even in analyses of a survey data set much larger than most, it appears that deletion and mean substitution methods do have some impact on the estimation of the model.

In Table 1, the use of the dummy variables in the fourth specification demonstrates that persons not responding to the education item, answering that they do not know their family income, not answering the governmental salience question or not responding to one or more of the items in the political efficacy scale are less likely to have voted than persons responding to these items. In general, the analysis suggests that persons not responding to political-attitude items are less likely to participate in the electoral process through voting. With respect to the political efficacy scale, the fact that a respondent did not respond to one or more of the efficacy items is more meaningful in the analysis than any set of valid responses. Respondents not reporting their age were more likely to have voted than respondents answering the age question.

When comparing the estimators for each set of valid and missing variables in the fourth specification, the analysis is consistent with the suppositions that older persons are less likely to respond to age items, that persons with less education and whose families have less income are less likely to respond to education and income questions, and that people who feel that government has little

TABLE 1

MODELS OF VOTER PARTICIPATION USING DIFFERENT SPECIFICATIONS FOR MISSING DATA

VARIABLES	CASE-WISE DELETION		PAIR-WISE DELETION		MEAN SUBSTITUTION		MEAN SUBSTITUTION AND DUMMY VARIABLES	
	ESTIMATOR	STANDARD ERROR OF ESTIMATOR	ESTIMATOR	STANDARD ERROR OF ESTIMATOR	ESTIMATOR	STANDARD ERROR OF ESTIMATOR	ESTIMATOR	STANDARD ERROR OF ESTIMATOR
EDUCATION (YEARS)	.033	.0025	.038	.0025	.039	.0017	.036	.0017
AGE (10 YEAR UNITS)	.032	.0048	.041	.0046	.038	.0032	.041	.0032
FAMILY INCOME ($10,000 UNITS)	.018	.0055	.029	.0059	.021	.0050	.023	.0050
RACE (WHITE)	-.039	.0160	-.051	.0160	-.032	.0110	-.043	.0110
POLITICAL EFFICACY SCALE	-.0028	.0160	-.0021	.0048	-.0016	.0036	-.001	.0036
POLITICAL INTEREST SCALE	.004	.00081	.0061	.00089	.0066	.00062	.0058	.00061
GOVERNMENTAL SALIENCE	.029	.0150	.040	.0150	.043	.0110	.044	.0100
CONSTANT	.21		.047		.0071		.098	
MISSING DATA INDICATORS:								
MISSING EDUCATION							-.190	.0240
MISSING AGE							.130	.0300
MISSING FAMILY INCOME							-.045	.0190
'DON'T KNOW' FAMILY INCOME							-.097	.0110
REFUSED TO GIVE FAMILY INCOME							.018	.0160
MISSING RACE							.025	.0250
MISSING POLITICAL EFFICACY							-.058	.0130
MISSING POLITICAL INTEREST							-.028	.0180
MISSING GOVERNMENTAL SALIENCE							-.090	.0190
R^2	.13		.16		.14		.17	
STANDARD ERROR	.36		.41		.41		.40	
N	2646		3138		7018		7018	

impact on them are less likely to respond to the governmental salience question. It further suggests that the estimators for the valid education, age, income, political efficacy and governmental salience variables are rendered slightly more conservative by the occurrence of missing observations. Thus the analysis using the mean substitution and dummy variable specification permits the researcher to assess the systematic effect introduced by missing observations.

While the use of the mean substitution and dummy variable specification provides more information for use in interpretation, and while the bias attributable to missing observations is discernible, such bias, in the analysis presented here, does not appear to be such that it would lead to misinterpretation when commonly used missing data routines are employed. All of the specifications included in Table 1 support the hypothesis that higher socio-economic status persons are more likely to vote. While political interest and feeling that government affects one as an individual (governmental salience) are positively related to voting, political efficacy seems to have little, if any, effect on voting. "Being White," when other variables are controlled, has a negative effect on voting. This finding is inconsistent with some prior voting behavior studies, but is supported by previous research in Atlanta, a city where Black candidates are major actors in electoral politics (Collins, 1977). The systematic bias introduced by missing observations in this model seems to have moderate impact; some bias is apparent and such bias could be much more severe when analyzing smaller data sets or data sets with more missing observations. In such instances, bias may lead to misinterpretation if the assumption of random occurrence of missing observations is not investigated.

Even though there were few missing observations (.8%) in the dependent variable in this example, the use of discriminant analysis will illustrate the utility of the mean substitution and dummy variable procedure in estimating the effects of missing observations in dependent variables. Discriminant analyses were performed on two specifications of the voter participation model; the first employed voting, non-voting and missing categories and the mean substitution procedure to "plug" missing observations in the discriminating variables. In the second discriminant analysis, the same categories were used, but the mean substitution and dummy variable procedure was used to create a missing index for each discriminating variable and these indices were employed along with the valid discriminating variables.

In the first analysis, using the mean substitution procedure, 57.03% of the "grouped" cases were correctly classified. When employing the mean substitution and dummy variable specification, 65.5% of the "grouped" cases were correctly classified. It is clear, then, that including indices of missing observations provides additional substantive information. Again, this procedure, unlike most commonly used procedures that attempt to "get rid" of missing data,[5] allows the researcher to determine if the fact that missing data occurred has substantive import. In this case, it obviously does.

Additionally, employing commonly used missing data routines without investigating the random occurrence assumption can increase rather than decrease existing bias attributable to missing observations. Table 2 presents the prediction table resulting from the discriminant analysis employing the mean substitution and dummy variable specification. Note that missing observations did not occur randomly; the analysis shows that the missing category could be systematically predicted along with the substantive categories. Furthermore, this analysis suggests that respondents that did not answer the voting question were more likely, if they had responded, to have reported that they did not vote than to have reported that they did vote. Seventy-three percent of the subjects responding to the voting item reported that they had voted ($\bar{X}$=.73). Thus, the analysis suggests that had the mean substitution procedure been employed to "plug" or "get rid" of missing observations in the voting participation variable, bias would have been exacerbated rather than mitigated. If the missing observations in the voting variable were to be "plugged," the value assigned them should reflect the greater likelihood of the respondents having reported that they did not vote. If cases with missing observations in the voting variable were to be allocated either to voting or non-voting categories, it would be appropriate, on the basis of the variables employed here, to allocate a larger proportion of these cases to the non-voting category.

Summary

In the example above, it is shown that the assumption of random occurrence of missing observations may be investigated in order to determine if bias is introduced by missing data. If bias is identified, the procedures explained above can help the researcher in understanding the nature of such bias, thereby enhancing interpretation of the analysis. While bias attributable to the occurrence of missing observations may or may not influence analysis in a manner which affects interpretation, we contend that the mean substitution

TABLE 2

PREDICTION RESULTS OF DISCRIMINANT ANALYSIS USING VOTED, DID NOT VOTE AND "MISSING" VOTING CATEGORIES AND EMPLOYING VALID AND MISSING INDICES AS DISCRIMINATING VARIABLES

ACTUAL GROUP	PREDICTED GROUP			N
	VOTED	DID NOT VOTE	"MISSING" VOTING	
VOTED	[64.1%]	31.0%	4.8%	5136
DID NOT VOTE	23.4%	[69.7%]	6.9%	1882
"MISSING" VOTING	18.0%	27.9%	[54.1%]	61

% OF "GROUPED" CASES CORRECTLY CLASSIFIED: 65.5%

and dummy variable procedure explained above provides a relatively simple method for clarifying the effects of missing data and incorporating additional information into interpretation and analysis. Missing data, here, are viewed as additional information to be analyzed and understood, rather than discarded.

Commonly used missing data routines employed without investigating the random occurrence assumption have several disadvantages. These routines not only "get rid" of information that may be useful in the analysis of survey data, but they also, as has been demonstrated above, may obscure or even increase existing biases. The mean substitution and dummy variable procedure permits the researcher to use all available information in an attempt to fully understand the effects of both valid and missing observations.

Notes

1. See for examples: Berelson, Lazarsfeld and McPhee (1954); Campbell, Converse, Miller and Stokes (1964); and Verba and Nie (1972).

2. Education is defined as number of years of formal schooling; family income is defined as total family income, before taxes and other deductions, for the calendar year 1975. The political efficacy and interest variables are three-item scales adapted from Verba and Nie (1972: 367-370).

3. The numbers of missing observations for the independent and control variables in the model are as follows: Missing Education (342); Missing Age (224); Missing Race (266); Missing Political Efficacy (1,469); Missing Political Interest (562); Missing Governmental Salience (635); Missing Family Income (524); Refused to Give Family Income (772); "Don't Know" Family Income (1,963).

4. Standard errors of estimators in regressions employing dummy variables should be interpreted cautiously since there is a tendency for the standard errors to be artifically deflated (Matloff, 1977). Therefore, we have interpreted the stability of the estimators in this model conservatively.

5. For examples of such procedures in discriminant analysis, see Chan, Gilman and Dunn (1976).

References

Ben-Sira, Zeev (1977) "A Facet Theoretical Approach to Voting Behavior," Quality and Quantity 11 (June): 167-188.

Berelson, B.R., P.F. Lazarsfeld and W.N. McPhee (1954) Voting. Chicago: University of Chicago Press.

Campbell, A., P.F. Converse, W.E. Miller and D. E. Stokes (1964) The American Voter. New York: John Wiley and Sons.

Chan, L.S., J.A. Gilman and O.J. Dunn (1976) "Alternative Approaches to Missing Data in Discriminant Analysis." Journal of the American Statistical Association 71 (December): 842-844.

Cohen, Jacob (1968) "Multiple Regression as a General Data-Analytic System," Psychological Bulletin 70 (December): 426-443.

Cohen, Jacob and Patricia Cohen (1975) Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. New York: John Wiley and Sons.

Collins, William P. (1977) "Race as a Salient Factor in Non-Partisan Elections." Unpublished Manuscript, Department of Political Science, Georgia State University.

Gleason, Terry C. and Richard Staelin (1975) "A Proposal for Handling Missing Data." Psychometrika 40 (June): 229-252.

Hertel, Bradley R. (1976) "Minimizing Error Variance Introduced by Missing Data Routines in Survey Analysis." Sociological Methods and Research 4 (May): 459-474.

Hesseldenz, Jon S. (1976) "Determining Validity and Identifying Nonresponse Bias in a Survey Requesting Income Data." Research in Higher Education 5 (April): 179-191.

Matloff, N.S. (1977) "Realistic Standard Errors of Estimated Regression Coefficients." Paper presented at the Western Meeting of the Institute of Mathematical Statistics and the American Statistical Association, June 22-24, Stanford California.

Press, S.J. and A.J. Scott (1976) "Missing Variables in Bayesian Regression, II." Journal of the American Statistical Association 71 (June): 366-369.

Verba, Sidney and Norman H. Nie (1972) Participation in America. New York: Harper and Row.