IDENTIFICATION OF NONRANDOM ITEM RESPONSE IN AN EPIDEMIOLOGIC SURVEY OF THE ELDERLY

Dwight B. Brock, National Institute on Aging, Jon H. Lemke and Robert F. Woolson, University of Iowa

INTRODUCTION

One of the most common difficulties encountered when analyzing surveys is the problem of nonresponse, both unit and item nonresponse. Unit nonresponse occurs when sampled subjects do not participate in the study, while item nonresponse occurs when participants in the survey fail to provide answers to some of the questions. A variety of possible reasons exist for each type of nonresponse, and each of these can lead to different potential nonresponse biases.

In this report we describe an approach for examining item nonresponse, in which our major emphasis is to describe a procedure for identifying variables which are related to nonresponse. Traditionally, approaches to handling item nonresponse have included ignoring the cases with missing items, imputing values for the missing data, or using inferential techniques such as the E-M algorithm (Fuchs, 1982). All these approaches are based on the assumption that the data are missing at random, whereas the present approach attempts to relate certain measured characteristics to nonresponse. We develop our approaches to identify factors related to the types of incompleteness indicated for the dependent variables. By partitioning the missing data as "refusals," "unable to answer," or inadvertently missing, we gain insight into the factors related to the types of nonresponse. Our approach emphasizes the weighted least squares (WLS) approach to categorical data analysis (Grizzle, Starmer and Koch, 1969). The general thrust is to examine specific contrasts, and general linear functions to detect meaningful patterns of incomplete responses. Finally, we apply these techniques to analyze the item nonresponse patterns among participants for the cognitive recall test data from the Iowa Established Population for Epidemiologic Studies of the Elderly (EPESE) (NIH, 1986).

GSK APPROACH TO CATEGORICAL DATA

The Grizzle, Starmer and Koch (GSK, 1969) approach to categorical data analysis is a well-known WLS technique for the analysis of a general class of linear and log-linear models. The approach is particularly amenable to analyses of certain functions of the cell probabilities which are expressed in a linear model involving the various population effects. In the next section these techniques will be used for modeling various logistic functions of the incomplete data.

To fix ideas it is useful to establish notation. We express this notation in a form convenient for the problem of nonresponse data analysis. For simplicity, it is assumed that there is one response variable which has r_0 categories. In addition, it is assumed that there are s subpopulations which are defined by the population or factor variables of interest. Since not all individuals agree to provide the response variable, the notation is expanded to allow the possibility of an additional $(r-r_0)$ categories for the categories of nonresponse. For example, there may be three categories of nonresponse including: unable to participate, refusals, and inadvertently missing. The expected cell probabilities are represented as follows:

Let $\underline{\Pi}_i = (\pi_{i1}, \pi_{i2}, \pi_{i3}, \pi_{i4})$, i=1,...,s, be the probabilities of given, "unable," refusal and missing observations for the ith of s subpopulations. These unknown probabilities are estimated by $\underline{P}_i = (P_{i1}, P_{i2}, P_{i3}, P_{i4})$, the observed cell proportions. The covariance matrix of $\underline{P}_i$, denoted $V_i(\underline{\Pi}_i)$ can be estimated from the data under a product multinomial model when the subpopulations have been sampled independently. In the case of complex sample surveys $\underline{P}_i$ is a vector of weighted sample estimates including any post sampling adjustments in the weights, and the covariance matrix can be estimated using a variety of techniques such as balanced repeated replication (see, for example, Kish and Frankel, 1970) or Taylor series linearization methods (Woodruff, 1971). As long as consistent estimators of the covariance matrix can be found, the analysis can be carried out in the same way as for the multinomial case (see Freeman, et al., 1976 and Freeman and Brock, 1978).

If these probabilities are homogeneous across all the subpopulations, there is no information to improve on the standard assumption of missing at random. If the probabilities are not homogeneous, the usual tests of homogeneity may not apply because of difficulties of interpretation or sparseness of the tables. Both of these problems are often avoided by considering the following functions:

$$f_1(\underline{\Pi}_i) = \ln[\pi_{i1}/(1-\pi_{i1})] \quad (1)$$

$$f_2(\underline{\Pi}_i) = \ln[\pi_{i2}/(1-\pi_{i2})] \quad (2)$$

$$f_3(\underline{\Pi}_i) = \ln[\pi_{i3}/(1-\pi_{i3})] \quad (3)$$

$$f_4(\underline{\Pi}_i) = \ln[\pi_{i4}/(1-\pi_{i4})] \quad (4)$$

These response functions are the natural logarithms of the odds of each possible response versus the sum of the probabilities of the other possible responses.

In the application of the GSK procedures, attention is focused on a vector of functions $\underline{F}(\underline{\Pi})$. In the case of our example

$$\underline{F}(\underline{\Pi}) = [f_1(\underline{\Pi}), f_2(\underline{\Pi}), f_3(\underline{\Pi}), f_4(\underline{\Pi})] \quad (5)$$

from above. A consistent estimator of the covariance matrix of $\underline{F}(\underline{P})$ is

$$S = H V(\underline{P}) H'$$

where H is the matrix of partial derivatives of $f_k(\underline{\Pi})$ with respect to Π_{ij}, evaluated at $\Pi ij = Pij$. The GSK method then applies WLS techniques to linear models for $\underline{F}(\underline{P})$. In particular, for the model

$$\underline{F}(\underline{\Pi}) = X\underline{\beta} \tag{6}$$

the GSK estimator for β is

$$\hat{\underline{\beta}} = (X'S^{-1}X)^{-1}X'S^{-1}\underline{F}(\underline{P}). \tag{7}$$

A test of fit of the model is given by

$$(\underline{F}(\underline{P}) - X\hat{\underline{\beta}})' S^{-1} (\underline{F}(\underline{P}) - X\hat{\underline{\beta}}), \tag{8}$$

which is asymptotically chi-square with degrees of freedom equal to the number of rows minus the number of columns of X if the model holds (Wald, 1943). Tests of specific linear hypotheses of the form $H_o : C\underline{\beta} = \underline{0}$ may be conducted using the statistic

$$(C\hat{\underline{\beta}})' (C(X'S^{-1}X)^{-1} C\hat{\underline{\beta}} \tag{9}$$

which is chi-square with degrees of freedom equal to the row rank of C if H_o is true.

GSK METHODS FOR IDENTIFICATION OF FACTORS INFLUENCING NONRESPONSE

In this section certain functions of the item nonresponse categories which fit within the GSK framework are studied. It is assumed that the objective is to model item nonresponse of the participants as a function of the factors which identify the subpopulations. The response functions are defined within subpopulations and contrasts across subpopulations to establish the nonhomogeneity of the response functions. Grizzle (1971) has illustrated the application of GSK techniques to the problem of multivariate logit analysis. In this setting, several logits are constructed for each subpopulation and these are modeled simultaneously as a function of the subpopulations of interest. For our purposes, it may be fruitful to construct a logit for complete response, a logit for "unable to respond," a logit for "refusals," and a logit for "missing (unspecified reason)" as in equations (1)-(4). We can model the four logits simultaneously by considering design matrices specifying subpopulations for the four respective logits. In this case the test of the fit (8) represents a test of the fit of the simultaneous models to the data.

The primary advantage to the simultaneous fitting of logits is the ability to contrast the logit model for one category of nonresponders to another category of nonresponders. Since the logits are in general correlated, it is important to incorporate this correlation into the statistical analyses. Individual logit analyses would not produce estimates of these correlations while the simultaneous analysis would. In the following section we apply this general framework for identifying subpopulations with different patterns of nonresponse.

EXAMPLE

The Established Populations for Epidemiologic Studies of the Elderly (EPESE) were designed to survey a broad range of physical, mental, and social characteristics of the elderly in four communities: East Boston, Massachusetts; New Haven, Connecticut; a five-county area surrounding Durham, North Carolina; and Iowa and Washington Counties in Iowa (NIH, 1986). These studies, sponsored by the National Institute on Aging, are longitudinal, consisting of comprehensive baseline interview surveys, followup interviews at yearly intervals and surveillance for hospitalization, nursing home utilization, and mortality end points subsequent to the baseline survey. The examples discussed in this paper are taken from the baseline survey conducted in Iowa and Washington Counties, Iowa, two rural communities in East Central Iowa. The target population for this site consisted of all persons 65 years or older living in the two counties as of December 1, 1981, the starting date for the Iowa baseline survey. The enumerated target population consisted of 4,601 individuals at least 65 years of age; 3,673 (80 percent) of these were interviewed. Nonrespondents consisted of 872 refusals, 36 too ill and without proxy, and 21 with no contact. Due to illness and other extenuating circumstances, 16 percent of the 80 percent studied received either abbreviated (120), telephone (250), or proxy interviews (206).

Since item nonresponse was anticipated when this study was planned, the questionnaire was designed with explicit codes for the various types of possible nonresponse. Specific categories of "Don't Know," "Refusals," "Unable to Answer," and simply "Missing" were included for the major variables. Thus, by considering the specific type of incompleteness, the unreasonable assumption that the data are missing at random need not be made when analyzing the data.

The cognitive recall function test was asked only of the subpopulation of 3,097 individuals who completed the full interview. The cognitive recall test consists of a count of the number of words the interviewee had correctly recalled from a list of 20 words. Eighty-eight (2.8 percent) of the 3,097 individuals were unable to take the cognitive recall test. On the other hand, 186 (6.0 percent) of the interviewees refused participation, while the data for 24 (0.8 percent) of the individuals were missing with no specific reason indicated.

To illustrate the methods of the last section, the specific categories of nonresponse are modeled here with the logistic function. These logits are then expressed as a function of age and sex, sex and education, sex and self-perceived health status, and sex

and prior coronary history. Other variables have been examined, but for brevity those results are omitted. Each of these four analyses is extremely informative, and describes the full range of application of the GSK methods described in the previous section. Unfortunately, the five variables cannot be modeled simultaneously since the resulting table is quite sparse.

In Table 1 item nonresponse for cognitive recall assessment is evaluated as a function of age and sex. It may be noted from this table that age has a significant impact on the "unable to participate" nonresponse category. In addition, it may be noted that an appreciable amount of this age effect is explained by a linear age effect. Indeed, examination of the "unable to participate" nonresponse rates suggests an increasing nonresponse rate with increasing age. This pattern holds for both males and females. In contrast, the analysis for the "refusals" reveals a significant sex effect and a significant age effect. In particular, males have higher refusal rates than females, while the age effect is largely accounted for by a linear component for females. Finally, the analysis of the "missing" nonresponse category shows only an overall age effect which is not accounted for by just a linear effect. No other effects are significant for the "missing" nonresponse logit when analyzing age and sex together.

Table 2 contains an analysis of sex and education on the type of nonresponse. Here it may be noted that education is predictive of nonresponse for both the "unable to participate" and the "refusals." On the other hand, education is not important for the "missing" nonresponse category. The education effects for the "unable to participate" and the "refusals" are in the direction of higher nonresponse rates for the lower educated than for the higher educated interviewees. The sex effects are not significant in the presence of education.

Table 3 is an analysis in which sex and self-perceived health status are considered jointly in their ability to predict the type of nonresponse. It may be noted that there is a significant nonlinear effect of self-perceived health status on the "unable to participate." In contrast, the self-perceived health status effect on the "refusals" is largely accounted for by a linear effect, with the higher refusal rates being associated with the poor self-perceived health status category, and the lowest refusal rates associated with the excellent self-perceived health status category.

In Table 4, an analysis is performed by the sex and the coronary medical history of the interviewee. In this case, there is an interaction of the two for the "refusal" nonresponse category. In particular, males with no history of a heart attack have a higher rate of refusal than males with history of a heart attack; the reverse is true for the females.

SUMMARY AND CONCLUSIONS

To summarize all of these analyses, the "unable to participate" in the recall test are related to the age, the education, and the self-perceived health status of the interviewee. In contrast to this, the "refused" nonresponse category is related not only to these three variables but also the sex and the interaction of sex and the coronary history of the interviewee. Finally, the "missing" nonresponse category is related to the age of the interviewee. Five variables, therefore, are related to nonresponse in the analysis. The attempt to minimize nonresponse in this survey does not allow us to analyze all five variables simultaneously; thus, one cannot argue for parsimony when identifying imputation classes.

Significantly, the analyses presented here indicate that proper subsets of these five variables are related to those "unable to participate" and those with "missing." Imputation for nonresponse would be done in this situation by considering the type of nonresponse recorded on the interview. Individuals who were "unable to participate" should have their cognitive recall scores imputed by considering only the age, education, and self-perceived health status. An individual who is "missing" would have an imputation performed by consideration of his/her coronary history and age. Finally, imputation for "refusals" would be done using a wide range of characteristics including sex, age, education, self-perceived health status, and coronary history.

In this work, we approach the problem of item nonresponse with the realization that missing data are often not missing at random. While rejecting this "missing-at-random" assumption, we presume that there are subpopulations for which this assumption can legitimately be made. Naturally, we can do little but assume randomness if the nonresponse is independent of the measured subpopulation variables.

When one ignores subjects with missing data or fails to recognize factors affecting the nonresponse, biases may be introduced. For instance, an individual who is very ill may have a likelihood of responding quite differently from someone who is too busy and believes interviews are a waste of time. The manner in which the biases influence the analysis depends on the purpose of a specific analysis, e.g., are we estimating the prevalence of a condition in the population, or are we investigating the association of several variables?

As we have seen, WLS analyses for categorical data are readily adapted to identify factors related to the probabilities of item nonresponse, through their ability to

provide simultaneous modeling of multiple linear compounds of the data. The primary importance of these methods lies in the influence the nonresponse biases would have on other analyses of the database for tests of the principal study hypotheses.

REFERENCES

Freeman, D.H. and Brock, D.B. (1978). The role of covariance matrix estimation in the analysis of complex sample survey data. In Namboodiri, N.K., ed., Survey Sampling and Measurement. New York: Academic Press, Inc.

Freeman, D.H., Freeman, J.L., Brock, D.B., and Koch, G.G. (1976). Strategies in the multivariate analysis of data from complex surveys. II: an application to the United States National Health Interview Survey. International Statistical Review 44, 317-330.

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

Grizzle, J.E., Starmer, C.F., and Koch, G.G. (1969). Analysis of categorical data by linear models. Biometrics 25, 489-504.

Grizzle, J.E. (1971). Multivariate logit analysis. Biometrics 27, 1057-1062.

Kish, L. and Frankel, M.R. (1970). Balanced repeated replication for standard errors. Journal of the American Statistical Association 65, 1071-1094.

National Institutes of Health (1986). Established Populations for Epidemiologic Studies of the Elderly: Resource Data Book. NIH Publication No. 86-2443. Washington, D.C.:U.S. Government Printing Office.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society 54, 426-482.

Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate. Journal of the American Statistical Association 66, 411-414.

ACKNOWLEDGMENTS

The research of Jon H. Lemke and Robert F. Woolson was supported in part by NIA contract NO1-AG-0-2106 and NCI Grant 1 R01CA39065. The authors wish to acknowledge the work of Greg Drube in statistical computing and Karen Hewitt in clerical support.

TABLE 1. COGNITIVE RECALL RESPONSE BY SEX AND AGE

		TYPE OF RESPONSE			
SEX	AGE	GIVEN	UNABLE	REFUSAL	MISSING
MALE	65-69		1.9%	7.2%	0.6%
	70-74		0.9%	6.1%	0.9%
	75-79		4.0%	6.0%	2.4%
	80+		8.5%	8.9%	0.5%
FEMALE	65-69		1.4%	2.5%	0.4%
	70-74		0.8%	4.7%	0.2%
	75-79		2.5%	6.8%	1.1%
	80+		5.8%	8.1%	0.8%
SEX				(0.0400)	
AGE			(0.0001)	(0.0100)	(0.0500)
LINEAR			(0.0001)	(0.0001)	
SEX BY AGE				(0.0600)	
LINEAR (M)			(0.0001)		
LINEAR (F)			(0.0001)	(0.0001)	

TABLE 2. COGNITIVE RECALL RESPONSE BY SEX AND EDUCATION

		TYPE OF RESPONSE			
SEX	EDUCATION	GIVEN	UNABLE	REFUSAL	MISSING
MALE	<12 YEARS		4.4%	7.7%	
	≥12 YEARS		1.7%	5.6%	
FEMALE	<12 YEARS		3.4%	6.8%	
	≥12 YEARS		1.8%	4.2%	
SEX					
EDUCATION			(0.0001)	(0.010)	
SEX BY EDUCATION					

TABLE 3. COGNITIVE RECALL RESPONSE BY SEX AND SELF-PERCEIVED HEALTH

		TYPE OF RESPONSE			
SEX	HEALTH STATUS	GIVEN	UNABLE	REFUSAL	MISSING
MALE	EXCELLENT		2.2%	5.3%	0.4%
	GOOD		2.8%	6.6%	0.8%
	FAIR		4.7%	6.7%	1.6%
	POOR		4.5%	13.6%	1.5%
FEMALE	EXCELLENT		1.3%	4.6%	0.3%
	GOOD		2.1%	4.8%	0.6%
	FAIR		4.8%	6.4%	0.7%
	POOR		2.2%	10.1%	2.2%
SEX					
			(.0100)	(.0200)	
HEALTH STATUS				(.0001)	(.0500)
LINEAR					
SEX BY HEALTH				(.0300)	
LINEAR (M)				(.0300)	(.0700)
LINEAR (F)					

TABLE 4. COGNITIVE RECALL RESPONSE BY SEX AND HISTORY OF HEART ATTACK

		TYPE OF RESPONSE			
SEX	HEART ATTACK	GIVEN	UNABLE	REFUSAL	MISSING
MALE	NO		2.9%	7.4%	0.8%
	YES		4.7%	5.4%	1.6%
FEMALE	NO		2.5%	5.0%	0.7%
	YES		3.2%	9.1%	0.0%
SEX					
HEART ATTACK					
SEX BY HEART ATTACK				(0.0200)	