

LATENT CLASS MODELS FOR CONTINGENCY TABLES WITH MISSING DATA

Christopher Winship
Harvard University

Robert D. Mare

UCLA

John Robert Warren

University of Washington

September 1999

This research was supported by National Science Foundation Grants SBR-94-11875 and SBR-94-11670, "Loglinear and Logit Models of Structural Effects: Selection, Endogenous Treatment, and Choice" and by the Graduate School of the University of Wisconsin-Madison.

INTRODUCTION

Missing data is a common problem in many types of data analysis. In this paper we show how to deal with missing data in loglinear analyses of frequency tables.¹ Our approach is based on two ideas: (1) that latent class models can be adapted to contingency tables with missing data by defining variables that are latent (missing) for some cases and are manifest (observed) for others; and (2) that a latent class models can be viewed as loglinear models for tables in which some cells are unobserved or partially observed. Using our approach, we can retain the loglinear model framework and notation and deal with missing data through a modest extension of the standard model. Flexible software for latent class models, such as DNEWTON (Haberman 1989) and LEM (Vermunt 1996) is required, but the conceptual extension of elementary loglinear models is straightforward.² The analyses presented in this paper were done using an interaction version of Dnewton which is described in Appendix A. By explicitly incorporating missing data into the analysis of a contingency table, one can address two concerns. First, a researcher may be worried about the possible loss of statistical power or precision of estimation that results when observations with missing data are excluded from an analysis. If many cases have missing data on at least one variable, exclusion of these cases from the analysis may substantially reduce the sample and create unacceptably large standard errors. One may want to incorporate cases with missing data into the analysis so that the information associated with these cases can be used to obtain more precise estimates. When the loss of statistical power is the only problem, incorporating missing data into a loglinear analysis is usually straightforward.

Second, one may be concerned that exclusion of missing data may result in inconsistent parameter estimates in loglinear models. This is likely to occur if there is a systematic mechanism producing the missing data. In this case one should develop a model of the missing data process jointly with the substantive model of interest. Below, we discuss more precisely the types of missing data processes that lead to inconsistent estimates. Although correcting problems of this type can be difficult, it may be essential if one is to make appropriate substantive conclusions.. The next section of the paper presents an example that illustrates that how one deals with missing data affects both the precision of parameter estimates and in the substantive conclusions that one draws. The subsequent section briefly discusses one conventional approach to dealing with missing data, namely adding a category to a variable for missing data. We show that this procedure typically leads to inconsistent estimates. Next we show how a contingency table can be extended to incorporate missing data so that loglinear models for partially observed data can be applied. We then consider alternative assumptions about missing data and the models that these assumptions imply. We examine different models for our first example and then present a more complex empirical example. We then discuss various problems in estimation and identification. EXAMPLE 1 - PRENATAL CARE AND INFANT MORTALITY Panel a of Table 1 presents data on the relationship between prenatal care and infant mortality in two clinics. These data were first analyzed in Bishop, Fienberg, and Holland (1975). Little and Rubin (1987) supplement these data with the hypothetical data in Panel b of the Table, which contain 255 infants whose clinic ID is missing. A researcher who wishes to analyze the combined data in the two panels faces two problems. First, data are missing data for 255 out of 970 cases. If all cases with missing data were omitted, this would substantially reduce statistical power. This is a particularly serious issue because the response variable, infant mortality, measures a rare event.

Second, assumptions about the true values of the missing data may markedly affect the estimated effect of prenatal care on infant mortality. Table 2 illustrates a range of possible outcomes under various extreme assumptions. All other possible assignments of the missing data are less extreme in that they yield estimates that fall within the range of those reported in Table 2.

Estimates of the effect of prenatal care vary across a wide range. If clinic status is ignored (assumption 1) the estimated difference in mortality rates by level of prenatal care is large ($5.4\% - 2.6\% = 2.8\%$). However, if only complete data are used and estimates are computed within clinics (2), the estimated differences are very small (.4% for Clinic A and .1% for Clinic B). Assuming that all missing data are from Clinic A (3) or Clinic B (4) produces somewhat higher estimates of the effect of the difference (.5% and 1.4%

follow the form for likelihood terms in a log-linear model (e.g., Agresti 1990, p. 166), whereas the terms in the second summation follow the form for likelihood terms in a latent class model (Andersen 1980, p 260). Our missing data model, therefore, combines elements of log-linear and latent class models.

TYPES OF MISSING DATA MODELS

The discussion of missing data patterns in Table 2 showed that estimates of the effect of prenatal care depend critically on assumptions about how the missing data are distributed. Having shown how to include missing data into a loglinear analysis through the use of the expanded table, we can examine specific models. MCAR Models

The most restrictive model assumes that data are missing completely at random (MCAR) (Little and Rubin 1987). This means that the missing indicators are assumed to be independent of all other variables. Consider the hypothetical data in Table 3 where there are three variables, X, Y, and Z and there is missing information on Y for some cases. Let M be the missing data indicator for Y. Then the MCAR model for these data is equivalent to (M) (XYZ) where we have used the parentheses to indicate an arbitrary set of associations among X, Y, and Z. The models (M) (XY) (YZ), (M) (XZ) (Y), and (M) (XYZ) are all examples of models where the data are assumed to be MCAR, that is missing completely at random. In each of these cases whether data are missing, as indicated by M, is assumed to be independent of the other variables in the model. When data are MCAR, they can be omitted from the analysis without affecting parameter estimates. However, because omitting missing data reduces the sample size, it reduces the precision of estimated parameters involving variables that have no missing data. MAR Models

A less restrictive and usually more realistic assumption is that whether a variable is missing is a function of the values of other observed variables; that is, data are missing at random (MAR), conditional on the observed data (Little and Rubin 1987). With MAR data, whether a variable is missing for a particular case is random conditional on the observed values on the other variables. In terms of our hypothetical data in Table 3, MAR models are of the form (MXZ) (XYZ), where, as before, the associations among the variables inside the parentheses are arbitrary. These models are MAR because M is conditionally independent of the variable it pertains to, Y. Typically, the inclusion of MAR data has very small effects on the estimated relationships between the variable that has missing data and the other variables. However, dropping cases with missing data may affect the estimated associations among the other variables in the model. Thus incorporating cases with missing data into the analysis may affect both the precision and consistency of the estimates for the relationships among variables that do not have missing data (Winship and Mare 1989). NINR Models

When the probability that a variable has missing data is associated with the variable itself conditional on the other variables in the model, this requires a model for nonignorable nonresponse

or NINR. NINR models are required when it is likely that some survey respondents refuse to answer a question.. For example persons with unusually high or low incomes may be less willing to divulge their incomes than persons with incomes in the middle of the income distribution. Elsewhere (Winship and Mare 1989) we examined data on whether individuals had ever been arrested. Here, one would expect that persons who had been arrested would be less likely to answer the question than those who had not been arrested.⁴ In our hypothetical data, any model that assumes dependence between M and Y is a NINR model. For example (MY) (XY) (YZ), (MY) (MX) (XY) (YZ), and (MYZ) (XY) (YZ) are all NINR models. When data are in fact subject to nonignorable nonresponse, omitting observations with missing values results in estimates that are not only inefficient but inconsistent as well. Often the bias can be considerable. For example, in Table 2, cases 3 through 10 are consistent with a variety of NINR models. Depending on what assumptions one makes about the true values of the missing data, one gets quite different estimates for the effect of Prenatal Care on Survival. Unfortunately, NINR models can often be difficult to estimate.

FITTING MISSING DATA MODELS TO INFANT MORTALITY DATA We can illustrate the fitting of alternative missing data models with the data in Table 1. Table 4 presents the estimated likelihood ratio G² and BIC (Raftery 1995) statistics for the complete data and selected missing data models. In this analysis we regard infant mortality as the response variable and clinic and prenatal care as the explanatory variables. Thus, we only consider models that include the association between prenatal care and clinic (PC). The G² values here should be viewed with caution. Because several of the frequencies in Table 1 are very small, the G² statistics may not follow a χ^2 distribution (Agresti 1990).

For the complete data, a model specifying all one way effects and the two ways associations between Survival Status and Clinic, (SC), and between Prenatal Care and Clinic (PC), fits the data extremely well (G² = .08, df = 1), implying that Prenatal Care and Survival Status are conditionally independent within clinics. The data, therefore, suggest that the level of prenatal care does not affect the probability that an infant will survive (Bishop, Fienberg, and Holland 1975; Little and Rubin 1987). If we include the missing data, then a much larger class of models is available. Among MCAR models, the (M) (SC) (PC) model, which is analogous to the best fitting complete data model, fits the data well (G² = 7.98, df = 5). This is an adequate fit, assuming that G² does in fact follow a χ^2 distribution for these data. Although the fit of the MCAR model is reasonable, it is not nearly as good as the fit of the analogous

model to the complete data alone. Thus, it is of interest to examine MAR and NINR models for these data. Considering these models as a whole, one can draw a number of general conclusions. First, models that do not contain both the (SC) and (PC) associations fit the data poorly. Second, for all three types of missing data models, the (SP) association appears to be statistically insignificant. Thus, for these data, our analysis is consistent with the analysis of the complete data alone. The data support the assumption that prenatal care and survival status are conditionally independent.

Across the three types of missing data models, however, there are some interesting differences. MAR models fit better than their corresponding MCAR models. Whereas the (M) (SC) (PC) model has a G2 of 7.98 on 5 degrees of freedom, the two comparable MAR models where M is assumed to be a function of Level of Prenatal Care (8) or Survival Status (11) have G2's of 3.30 and 5.79 respectively (df = 4). Differencing these amounts from the MCAR G2 we get G2 statistics of 4.68 and 2.19 (df = 1). Assuming that G2 follows a χ^2 distribution, the first of these differences is significant at the .05 level, indicating that a significant improvement in fit is achieved by assuming that M is associated with level of prenatal care (P). Adding the (MS) term to the (MP) (SC) (PC) model lowers the G2 to 1.57 (Model 15) but this decrement is not statistically significant. The NINR models also fit the data better. A model in which the likelihood of having missing data is simply a function of the clinic, that is, (MC) (SC) (PC), gives a G2 of 2.26 on 4 degrees of freedom (Model 19). This model has a lower G2 with the same degrees of freedom than the MAR model (MP) (SC) (PC) (8), although a nested comparison between these two models is obviously impossible. The (MC) (SC) (PC) model implies that the likelihood of missing data on clinic is only associated with which clinic a mother was served by. Adding an (MS) term to the model lowers the G2 to 1.64 (df = 3) (27). Alternatively, adding an (MP) term to the (MC) (SC) (PC) model lowers the G2 to .39 (24), an extremely good fit. Neither of the changes, however, is significant at the .05 level, assuming a χ^2 distribution for G2. Based on considerations of fit, parsimony, and plausibility, we conclude that Model 19 is the most satisfactory model for the data.

It is certainly possible to conceive of more complex NINR models. Table 2 illustrated how various assumptions about the pattern of missing data lead to substantial differences in estimates for the effect of Level of Prenatal Care on Survival. Cases 5 through 10 in Table 2 are all equivalent to saturated NINR models, that is, models that fit the data perfectly and that have zero degrees of freedom. Many of these models are also unidentified and, by definition, cannot be tested against the data. All of these models assume that the missing mechanism has a three way or higher interaction with the clinic and another variable. For example, cases 5 and 6 assume a

three-way interaction among M, C, and P. Cases 9 and 10 assume a four-way interaction among M, C, P, and S. These models all assume that in the two clinics different types of data are likely to produce missing data on clinic. Without some assumptions we cannot rule out the estimates associated with these allocations. If, however, we assume that the missing data mechanism is the same in the two clinics except for the rate at which missing data occurs, then this rules out these possibilities. The (MC) (SC) (PC) model represents this assumption. EXAMPLE 2 -- INTERGENERATIONAL EDUCATIONAL MOBILITY

A common use for loglinear models is the analysis of intergenerational social mobility. These analyses typically focus on cross classifications of parents' and offsprings' socioeconomic characteristics derived from retrospective reports by the offspring. A problem for these analyses is missing data on the parents' characteristics. In this example, we examine the association between father's schooling and offspring's educational attainment, using data from the 1994 General Social Survey (GSS) and the Survey of American Families (SAF). The 1994 GSS, a cross section survey of the U.S. population, included a module on the socioeconomic characteristics of persons related to GSS respondents, including parents, children, spouses, and siblings. The SAF was a telephone survey administered to one randomly selected sibling of the GSS respondents (Mare and Hauser 1994). Table 5 cross classifies father's and respondent's schooling as reported by respondents in the GSS. These data were restricted to persons aged 18 and over who had a sibling who was interviewed in the SAF. Of the 836 persons included in the table, 73 or 8.7 percent failed to report their father's educational attainment. This may be construed as a relatively modest amount of missing data, and some researchers would simply omit observations in which fathers schooling is missing. This decision, however, may have a big effect on one's estimates and inferences. One way of seeing the consequences of omitting missing data is to examine the estimated distribution of father's schooling and association between father's and offspring's schooling under several alternative hypothetical patterns of missing data. Table 6 shows the four local odds ratios for the 3 x 3 table of father's schooling by offspring's schooling under four scenarios: (1) data missing completely at random; (2) missing observations on father's schooling are all drawn from respondents whose fathers have less than 12 years of schooling; (3) missing observations are all drawn from respondents whose father's have more than 12 years of schooling; and (4) missing observations are all drawn from the same level of schooling as the respondent. Under these alternative assumptions, the proportions of persons whose fathers have less than 12 years of schooling ranges from about 33 to about 40 percent. The estimated local odds ratios under alternative assumptions about missing data vary substantially. For example, the local cross

product ratio between whether a father is a high school graduate vs. a dropout and the corresponding contrast for offspring varies between 2.12 for data missing at random at 4.56 for data missing exclusively from persons who have the same schooling level as their fathers.

We can examine the association between father's and offspring's schooling while taking account of missing data by using variants of the missing data models discussed above. Although data may be missing completely at random, it is more likely that whether data are missing on father's educational attainment is associated with a respondent's own educational attainment and possibly father's educational attainment itself. The former problem may arise because better educated respondents are more likely to be cooperative and conscientious in answering survey questions. The latter problem may arise if respondents perceive that it is desirable to have a better educated father or if individuals with more poorly educated fathers are less likely to know their father's schooling. Some offspring of fathers with low levels of educational attainment may misreport their father's schooling, but others may simply not report it. By itself, Table 5 provides limited information with which to investigate the impact of missing data on our estimates of the distribution of father's education or of the association between father's and offspring's education. That better educated respondents may be more likely to report their father's schooling can be investigated with this table. If this is the only systematic source of missing data and if father's and offspring's educational attainments are associated, this idea can be represented as a saturated MAR model for this table. However, the idea that whether data are missing on father's schooling is associated with father's schooling itself requires a model of nonignorable nonresponse (NINR), which is not identified from these data. To investigate nonignorable nonresponse, requires a variable that is associated with father's schooling but not with whether father's schooling is missing. Such variables are difficult to find because most characteristics of an individual that are associated with father's schooling are also associated with the individual's propensity to report father's schooling. A solution to this problem is to use the responses to the same item in an independent interview conducted with a person related to the original respondent. The SAF asked a sibling of each GSS respondent to report on father's schooling. Table 7 cross classifies GSS respondent's report of father's schooling, GSS respondent's report of his or her own schooling, and the SAF respondent's - that is, GSS respondent's sibling's - report of father's schooling. This table includes categories for missing data on both reports of father's educational attainment and can be used to examine a variety of models for missing data.

These models can be regarded as applying to an expanded table that

includes separate dimensions for the substantive variables of interest and for whether or not these variables are missing. The 3 x 3 x 3 x 2 x 2 expanded table has the following five dimensions: GSS respondent's educational attainment (O), GSS respondent's report of father's educational attainment (FG), SAF respondent's report of father's educational attainment (FS), whether FG is missing (MG), and whether FS is missing (MS). For example, from Table 7 we can identify GSS respondents who have missing data on father's schooling classified by their own schooling and their siblings' reports of father's schooling. We do not know the educational attainment of these individuals' fathers (although a large fraction of these individuals have the same father as their sibling and, for them, their father's education can be inferred). Thus, for a given level of own schooling and sibling's report of father's schooling, GSS respondents who have missing data on their own report of father's schooling are distributed in an unknown way across categories of their father's schooling. Our models for missing data are based on selected interactions among the five dimensions of this expanded table.

Many logically possible models may be fit to the frequencies in Table 7. We limit the range of possible models through the following substantive considerations. First, because of the well-known correlation between the socioeconomic positions of parents and offspring, father's and offspring's educational attainment are associated. Indeed, this association provides the substantive interest in this table. Inasmuch as the GSS and SAF respondent's report of father's educational attainment (FG and FS) apply to the same individual in most families, they are two reports of the same trait and thus both of these measures are associated with offspring's schooling. Thus, all models should include the FGO and the FSO associations.⁵ Second, inasmuch as most siblings share the same father, their reports of father's schooling are likely to be strongly associated. Thus, all of our models include the FGFS association. Third, siblings' propensities to fail to report father's schooling may be associated, either because of a shared reluctance to provide this information or a shared ignorance of their father's schooling. For most models, therefore, we include the MGMS association. Fourth, it is an empirical question whether or not data are missing on father's schooling is associated with respondent's own educational attainment and with father's schooling itself. Thus, we examine alternative models with and without the OMG, OMS, FGMG, and FSMS associations. Finally, we assume that whether a person reports father's schooling is conditionally independent of his or her sibling's reported level of father's schooling, given the association between each sibling's reported level of father's schooling. Thus, we assume the absence of the FGMS and the FSMG associations. These are the key restrictions for identifying NINR models for these data. Table 8 presents goodness of

fit statistics for selected models fit to the observed data in Table 7. Model 1 is an MCAR model in that it assumes conditional independence of whether data are missing on the two measures of father's schooling from any of the other dimensions of the table. This model includes parameters for the marginal distributions of whether or not data are missing on FG and FS (MG and MS respectively), but no association between whether or not data are missing on these two variables. As indicated by both the likelihood ratio G^2 and the BIC statistics (Raftery 1995), this model fits very poorly.

Models 2-5 are MAR models. Model 2 includes a parameter for the association between MG and MS and fits the data much better than Model 1. This suggests that GSS respondents and their siblings both fail to report their fathers' schooling at a much higher rate than one would expect if their rates of nonresponse were statistically independent. Common family circumstances may determine whether or not offspring know their father's educational attainment. Models 3, 4, and 5 incorporate parameters for the association between GSS respondent's schooling and whether GSS and SAF respondents' have missing data on father's schooling. Inclusion of both of these associations significantly improves the fit of the model. The OMG association implies that better educated respondents differ from more poorly educated respondents in their level of cooperation with the survey or their knowledge of their father's schooling. The OMS association may arise because the table does not include a dimension for SAF respondent's own educational attainment. Given a strong correlation between siblings' educational attainments, we observe an OMS association when the schooling of SAF respondents is not taken into account. Models 6-9 are NINR models that include associations between FG and FS on the one hand and MG and MS on the other. Model 6 includes the FGMG and FSMS associations, but excludes the associations between MG and MS and GSS respondent's educational attainment (0). This model fits much better than Model 2, the corresponding MAR model ($G - G = 69.9$, 4 df, $p < .001$), and provides provisional evidence of nonignorable nonresponse. Model 6, however, does not fit the data well. A more stringent test for NINR is to estimate the FGMG and FSMS associations in the presence of the OMG and OMS associations. Models 7 and 8 include the FGMG and FSMS associations respectively, as well as OMG and OMS. Model 7 fits marginally better than Model 5, the corresponding MAR model ($G - G = 5.3$, 2 df; $p = .071$), whereas Model 8 fits much better than Model 5 ($G - G = 23.0$, 2 df, $p < .001$). This provides strong evidence for NINR for FS and somewhat weaker evidence for FG. Model 9 includes the OMG, OMS, FGMG, and FSMS associations and fits significantly better than Model 5 as well as the three other NINR models. By the likelihood ratio criterion Model 9 is the best fitting model ($G = 20.3$, 18 df, $p = .316$), although the BIC criterion indicates that Model 8 is somewhat more satisfactory.

Table 9 presents estimates of two-way association parameters and partial odds ratios from Model 9 which reveal the systematic nature of missing data on father's educational attainment. First, the estimated partial odds ratio for the two missing data indicators (MGMS) indicates a positive association between siblings' propensities to fail to report father's educational attainment. The SAF respondent is almost twice as likely to have missing data on father's schooling if his or her sibling in the GSS fails to report father's schooling than if the sibling reports father's schooling. Second, more highly educated GSS respondents are much more likely to report their father's schooling than their more poorly educated counterparts. Relative to GSS respondents who were high school dropouts, for example, those who had more than a high school degree had odds of not reporting father's schooling that are only one fifth as great. Finally, missing data on father's educational attainment is associated with the level of father's education. For example, for SAF respondents whose fathers had more than a high school degree, the odds of missing data are only about seven percent of the odds for SAF respondents whose father's were high school dropouts. Both the parameter estimates and the goodness of fit tests provide more evidence of nonignorable nonresponse for SAF respondents' reports of father's schooling than for GSS respondents' reports. This may occur because GSS respondents' own educational attainments are controlled in our models, whereas SAF respondents' attainments are not.

ESTIMATION AND IDENTIFICATION

Estimation

The estimates reported in this paper were calculated using DNEWTON, a flexible program for the estimation of latent class models, including models with latent cells such as the ones discussed in this paper (Haberman 1988). A good alternative to DNEWTON is LEM (Vermunt 1996), which has similar capabilities. The development of several user friendly latent class programs has substantially reduced the burden of estimating latent class models. Although modern software makes the estimation of loglinear models for missing data feasible, several estimation problems nonetheless are common. These include: (1) failure of the program to converge; (2) estimates on the boundary of the parameter space; (3) cells with small expected frequencies (<5), and (4) poor model fit. These important issues, which are problems for latent class models more generally, are discussed in the introductory chapter of this volume.

Identification

Consider the hypothetical data shown in Table 3 consisting of variables X, Y, Z but collapsed over Z. If we add to this a variable M indicating whether information on Y is missing for an case we then have a 23 table of X by Y by M. This is shown in Table 10 both as a 2 x 3 subtable of the observed table and as a 23 expanded table that is partially latent. In a complete eight-cell table, a fully

saturated model has seven parameters (plus a grand mean). However, Table 10 has only six cells -- four for the X-Y complete data subtable and two for the categories of X among respondents with missing data on Y. At most, therefore, a model with five parameters is identified. If one fits a marginal parameter for each of the dimensions X, Y, and M, then, among hierarchical models, the most complex five-parameter model can include at most two two-way interactions. The potential models are: (1) (MX) (XY) (MAR Model); (2) (MY) (XY) (NINR Model); and (3) (MX) (MY), which is not identified.

To see that the (MX) (MY) model is not identified, note that in Table 10b one can observe the association between X and Y for the complete data. The (MX) (MY) model assumes that the partial X-Y association (conditional on M) is zero, which can be tested using the complete data. That we can test that the X-Y interaction is zero means that a parameter for this association is identified, irrespective of whether other parameters of the (MX) (MY) model are not identified. Such a test would use one degree of freedom, leaving only four degrees of freedom for estimating the five parameters of the (MX) (MY) model. Thus the (MX) (MY) model can not be identified. MAR models are usually identified, even if data are missing on several variables (Little and Rubin 1987: 171-194). A sufficient condition for identification of MAR models is that some observations are complete on all variables and that these observations represent all possible combinations of the variables in the model (Fuchs, 1982). General rules for the identification of NINR models have not yet been developed. Some guidance is available, however, from results on the identifiability of NINR models for two-way tables and from rules for identification of latent class models. Little and Rubin (1987: 238-239) summarize the identifiability of models for two-way (J x K) tables in which one dimension of the table has missing observations. Denote the dimensions of the table by X and Y. Let all observations be present for X, but some observations on Y be missing. Let a third variable, M, denote whether or not data are missing on Y. Among the several log-linear models that can be fit to the X, Y, M table, NINR models are those that include the M-Y interaction. The only NINR model that is potentially identifiable is (MY) (XY), that is, a model in which the fully observed variable is associated with the partially observed variable, but is independent of whether or not data are missing on the partially observed variable. This model is identified if $J \geq K$, that is, if the number of categories of the fully observed variable is at least as large as the number of categories of the partially observed variable. In a 2 x 2 table $J = K$ and the model is just identified.

These results suggest that NINR models are identified if: (1) for every partially observed variable, there exists a fully observed variable that is conditionally independent of the missing data

indicator; and (2) the number of categories of the partially observed variable does not exceed the number of categories of the fully observed variable. The fully observed variable plays a role in identifying the model that is analogous to an instrumental variable in a structural equation model and the condition is equivalent to assuming that parameter for the two-way interaction between the fully observed variable and the missing indicator is zero. This restriction can sometimes also be met by assuming that higher order interaction terms are zero. Baker and Laird (1988), for example, estimate a model in which two fully observed variables both affect the missing indicator, but the model is identified because the three-way interaction among the variables is assumed to be zero. An intuitive way of understanding the identifiability of some NINR models is to note that they are often similar to standard latent class models that are analogous to factor models (e.g., Goodman 1974). Consider the NINR (MC) (SC) (PC) model for Little-Rubin data. If C were missing on every observation, the model would be a standard latent class model in which C is a latent "factor" and M, S, and P are its observed "indicators." The model assumes that M, S, and P are conditionally independent given C. If C were missing on every case, though, the model would not be identified without either another indicator for C or additional restrictions on the parameters. In the pure latent variable case we observe only the two-way relationship between P and S which is insufficient for estimating the associations between both S and C and also P and C. But when data are missing for only some of the cases, we observe the relationship between S, P, and C in the complete data. Thus we can estimate both the S-C and P-C associations simply from the complete data. Only the relationship between M and C needs to be estimated indirectly. When C is partially observed, therefore, this NINR model is identified.

CONCLUSION

Conventional methods of dealing with missing data in multivariate models run serious risks. Omitting observations with missing data from an analysis certainly reduces sample size (and thus the precision of estimates) and, at worst, may lead to severe biases in parameter estimates. Simply incorporating missing data by adding categories for missing data to the observed variables generally results in inconsistent estimates. Fortunately, alternatives to the conventional approaches provide powerful methods for investigating the degree to which data are missing systematically and for carrying out appropriate substantive analyses. The approach illustrated in this paper is to extend standard analyses of categorical data by recognizing that, when some data are missing, the resulting contingency tables have cells that are partially observed. We suggest that one analyze such data using latent class loglinear models for tables that have a mixture of fully and partially

observed cells. This approach enables one to investigate hypotheses about the mechanisms by which missing data come about as well as the substantive relationships of interest. Inasmuch as these models are simply variants of standard loglinear models, one can incorporate missing data using well known procedures for specifying multiway interactions in contingency tables. Unlike standard loglinear models for fully observed data, however, these models typically require that the analyst incorporate additional data or make simplifying assumptions to identify the relationships between substantive variables of interest and indicators of whether or not data are missing.

REFERENCES

- Agresti, Alan. 1990. *Categorical Data Analysis*. New York. John Wiley.
- Andersen, Erling B. 1980. *Discrete Statistical Models with Social Science Applications*. Amsterdam: North-Holland.
- Baker, Stuart G., and Nan M. Laird. 1988. "Regression Analysis for Categorical Variables with Outcome Subject to Nonignorable Nonresponse." *Journal of the American Statistical Association* 83: 62-69.
- Baker, Stuart G., William F. Rosenberger, and Rebecca Dersimonian. 1992. "Closed-Form Estimates for Missing Counts in Two-Way Contingency Tables." *Statistics in Medicine*, 11: 643-657.
- Bishop, Yvonne M, Steven Fienberg, and Paul W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT Press.
- Clogg, Clifford C. 1988. "Latent Class Models for Measuring." Pp. 173-205 in R. Langeheine and J. Rost (eds.) *Latent Trait and Latent Class Models*. New York: Plenum.
- Davis, James A., and Tom W. Smith. 1986. *General Social Surveys, 1972-1986 (MRDF)*. NORC ed. Chicago: NORC. Distributed by Roper Public Opinion Research Center, New Haven, Conn.
- Dayton, C. Mitchell, and George B. Macready. 1980. "A Scaling Model with Response Errors and Intrinsically Unscalable Respondents." *Psychometrika* 45: 343-356.
- Fuchs, Camil. 1982. "Maximum Likelihood Estimation and Model Selection in Contingency Tables With Missing Data." *Journal of the American Statistical Association* 77: 270-278.

Goodman, Leo A. 1974. "The Analysis of Systems of Qualitative Variables When Some of the Variables are Unobservable. Part I: A Modified Path Analysis Approach." *American Journal of Sociology* 79: 1179-1259.

Haberman, Shelby J. 1989. "A Stabilized Newton-Raphson Algorithm for Log-linear Models for Frequency Tables Derived by Indirect Observation." Pp. 193-211 in C. C. Clogg (Ed.), *Sociological Methodology* 1988 (Vol. 18).

Hagenaars, Jacques. 1988. "Latent Structure Models with Direct Effects Between Indicators." *Sociological Methods and Research* 16: 379-405.

. 1993. *Loglinear Models with Latent Variables*. Newbury Park, CA: Sage Publications.

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

McCutcheon, Alan L. 1987. *Latent Class Analysis. Quantitative Applications in the Social Sciences Paper No. 64*. Beverly Hills, CA: Sage University Press.

Park, Taesung and Morton B. Brown. 1994. "Models for Categorical Data with Nonignorable Nonresponse." *Journal of the American Statistical Association* 89:44-52.

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." Pp. 111-163 in P. V. Marsden, (Ed.), *Sociological Methodology* 1995, (Vol. 25).

Vermunt, Jeroen K. 1996. *Log-Linear Event History Analysis*. Tilburg: Tilburg University Press.

Winship, Christopher, and Robert D. Mare. 1989. "Loglinear Models with Missing Data: A Latent Class Approach." *Sociological Methodology* 1989, Clifford C. Clogg, editor. Basil Blackwell: 331-368.

. 1992. "Models for Sample Selection Bias." *Annual Review of Sociology* 18: 327-50.

Appendix A

Interactive Dnewton
(iDnewton)

This program is a fully interactive version of Haberman's (1989) Dnewton program, a program for estimating latent class as well as standard loglinear models. It can be downloaded from <http://www.wjh.harvard.edu/sociology/faculty/winship>. The program allows a researcher to use Haberman's Dnewton in a fully flexible manner. In interactive mode it creates a batch file to be submit to Dnewton, Dnewton then process this files, and the output is returned. The program has a large number of features:

Data can be read from an ASACII file format or outputed from either STATA or SPSSX.

Models are specified using standard loglinear notation. Both hierarchical and nonhierarchical can be easily specified in this notation.

To run new models, all that needs to be done is to change the model specification statement. This makes running new models on the same dataset vastly easier than with the original version of Dnewton.

Categorical variables can be expanded into dummy variables using either an ANOVA (-1,1) or regression specificaiton (0,1).

Missing values can be defined for all variables at once or on a variable by variable basis

A category or categories of variables can be mapped into other categories. This option is particularly useful when a particular response indicates that an individual belongs to one of several categories, but which category is unknown.

Variables may be latent for the entire population or latent, but observed for different components of the population.

Default starting values for the model parameters are provided. The user, however, may specify alternative starting values.

Interactive help screens exist for all options.

Interactive Dnewton allows a user to fit models in Dnewton considerably faster than with the original program. In a period of thirty minutes a user can easily run ten to twenty different models, a process that would take at least several hours using the original Dnewton. Speed is basically achieved in three ways: (1) in order to run new models the only thing that needs to be changed in the model statement line; (2) an easy to use and flexible set of routines is available for recoding data; (3) data produced by STATA or SPSSX can be automatically read by the program.

FOOTNOTES

Table 1. Contingency Table with Partially Classified Observations

Clinic(C)	Survival (S)		Died	Lived
	Prenatal Care (P)			
(a) Completely Classified Cases				
Clinic A	Less		3	176
More		4	293	
Clinic B	Less		17	197
More		2	23	
(b) Partially Classified Cases (Clinic Missing)				
Less		10	150	
More		5	90	

Source: (a) Bishop, Fienberg and Holland (1975), table 2.4-2. (b) Artificial data from Little and Rubin (1987, Table 9.8, page 187).

Table 2. Mortality Rates for Data in Table 1 Under Alternative Assumptions about Missing Data

ASSUMPTION	Level of Prenatal Care			Difference
	Clinic	Less	More	
1. Observed Data Collapsed over Clinic	5.4	2.6	2.8	
2. Complete Data A	1.7	1.3	.4	
B	7.9	8.0	.1	Missing are all:
3. Clinic = A	3.8	2.3	.5	
4. Clinic = B	7.2	5.8	1.4	
5. If Care = Less, Clinic = A	3.8	1.3	2.5	
If Care = More, Clinic = B	7.9	5.8	2.1	
6. If Care = More, Clinic = A	1.7	2.3	-.6	
If Care = Less, Clinic = B	7.2	8.0	-.8	
7. If Survival = Died, Clinic = A	6.9	2.3	4.3	
If Survival = Lived, Clinic = B	6.4	1.8	4.6	

8. If Survival = Lived, Clinic = A A 1.3 1.0 .3

If Survival = Died, Clinic = B B 12.6 23.3 -10.7

9. If Care=Less & Survival = Died, A 6.9 1.0 5.9 or Care= More & Survival = Lived, Clinic =A

If Care=More & Survival = Died, B 11.8 23.3 -11.5 or Care= Less & Survival = Lived, Clinic = B

10. If Care=Less & Survival = Died, A .9 3.0 -2.1 or Care= More & Survival = Lived, Clinic =B

If Care=More & Survival = Died, B 12.0 1.7 10.3 or Care= Less & Survival = Lived, Clinic = A

Table 3. Hypothetical Data for Showing the Effects of Using a Missing Data Category

No Missing Data (N = 3600)

Marginal Associations

X = 0 X = 1

Z = 0	1000	800	1000
Z = 1	800	800	1000

Odds Ratio = 1.56

Conditional Associations

Y = 0 Y = 1

X = 0 X = 1 X = 0 X = 1

Z = 0	800	400	200	400
Z = 1	400	200	400	800

Odds Ratio = 1.0 Odds Ratio = 1.0

With Missing Data

Conditional Associations

Y = 0 Y = 1 Y = Missing

X = 0 X = 1 X = 0 X = 1 X = 0 X = 1

Z = 0	640	320	160	320	200	160
Z = 1	320	160	320	640	160	200

Odds Ratio = 1.0

Odds Ratio = 1.0

Odds Ratio = 1.56

Table 4. Goodness of Fit of Selected Models for Infant Mortality Data.

Model G2 df BIC

Complete Data

1. SC PC .08 2 -13.06

MCAR Models

2.	M	S	PC	24.81	7	-23.33	
3.	M	SP	PC	20.02	5	-14.37	
4.	M	SC	PC	7.98	5	-26.41	
5.	M	SC	PC	SP	7.84	4	-19.67

MAR Models

6.	MP	S	PC	20.13	5	-14.26		
7.	MP	PC	SP	15.33	4	-12.18		
8.	MP	SC	PC	3.30	4	-24.21		
9.	MP	SC	PC	SP	3.16	3	-17.47	
10.	MS	PC	SP	17.83	4	-9.68		
11.	MS	SC	PC	5.79	4	-21.72		
12.	MS	SC	PC	SP	5.65	3	-14.98	
13.	MS	MP	PC	17.94	4	-9.57		
14.	MS	MP	SP	PC	13.55	3	-7.08	
15.	MS	MP	SC	PC	1.57	3	-19.06	
16.	MS	MP	SC	PC	SP	1.38	2	-12.37

NINR Models

17.	MC	S	PC	20.13	5	-14.26		
18.	MC	PC	SP	15.33	4	-12.18		
19.	MC	SC	PC	2.26	4	-25.25		
20.	MC	SC	PC	SP	2.00	3	-18.63	
21.	MC	MP	S	PC	20.13	4	-7.38	
22.	MC	MP	PC	SP	15.33	3	-5.30	
23.	MC	MP	SC	PC	.39	3	-20.24	
24.	MC	MP	SC	PC	SP	.39	2	-13.36
25.	MC	MS	PC	17.94	4	-9.60		
26.	MC	MS	PC	SP	13.55	3	-7.08	
27.	MC	MS	SC	PC	1.64	3	-18.99	
28.	MC	MS	SC	PC	SP	1.40	2	-12.35

Table 5. Offspring's Educational Attainment by Father's Educational Attainment

Offspring's Schooling Father's
 Schooling <12 12 >12

<12 46 135 113

12 12 75 145

>12 4 30 206

Missing 23 29 18

Source: 1994 General Social Survey, Respondents Aged 18 and over
 with a sibling interviewed in the Survey of American Families

Table 6. Local Odds Ratios and Distributions of Father's Schooling
 Under Alternative Assumptions about Missing Data

Local Odds Ratios Distribution of Father's Schooling

<12/12 12/>12 <12 12 >12

1. Observed Data

<12/12	2.13	2.31			
12/>12	1.20	3.55	.384	.303	.313

2. Missing Data All Taken From Diagonal

<12/12	4.43	1.67			
12/>12	.87	5.36	.354	.293	.352

3. Missing Data All Taken From <12

<12/12	2.62	2.42			
12/>12	1.20	3.55	.409	.257	.332

4. Missing Data All Taken from >12

<12/12	2.13	2.31			
12/>12	.35	1.96	.328	.257	.413

Source: Table 5

Table 7. GSS Respondent's Educational Attainment by Father's
 Educational Attainment Reported by GSS Respondent by Father's
 Educational Attainment Reported by SAF Respondent

		Father's Schooling (SAF)										
		<12		12		>12		Missing				
Father's Schooling (GSS)	<12	>12	Offspring's Schooling <12	Offspring's Schooling 12	Offspring's Schooling >12	Offspring's Schooling <12	Offspring's Schooling 12	Offspring's Schooling >12	Offspring's Schooling <12	Offspring's Schooling 12	Offspring's Schooling >12	
<12	33	104	88	1	11	13	0	2	2	12	18	10
12	2	11	16	7	49	109	1	13	16	2	2	4
>12	0	0	3	1	9	23	3	21	177	0	0	3
Missing	12	17	5	4	6	6	0	3	2	7	3	5

Table 8. Goodness of Fit of Selected Models for Educational Mobility Table.

Model G2 d.f. BIC

1. FGFS, FGO, FSO, MG, MS (MCAR) 1470.6 27 1288.9
2. FGFS, FGO, FSO, MGMS (MAR) 117.9 26 -57.0
3. FGFS, FGO, FSO, OMG, MGMS, (MAR) 73.6 24 -87.9
4. FGFS, FGO, FSO, OMS, MGMS (MAR) 87.1 24 -74.4
5. FGFS, FGO, FSO, OMG, OMS, MGMS, (MAR) 51.5 22 -96.5
6. FGFS, FGO, FSO, FGMG, FSMS, MGMS, (NINR) 48.0 22 -100.0
7. FGFS, FGO, FSO, OMG, OMS, FGMG, MGMS (NINR) 46.2 20 -88.4
8. FGFS, FGO, FSO, OMG, OMS, FSMS, MGMS (NINR) 28.5 20 -106.1
9. FGFS, FGO, FSO, OMG, OMS, FGMG, FSMS, MGMS (NINR) 20.3 18 -100.8

Table 9. Estimated Association Parameters for NINR Model (Model 9)

Model Terms ? S.E.(?) exp(?)

MGMS .670 .358 1.954

FG12FS12		3.868	.294	47.847
FG12FS>12		3.967	.570	52.826
FG>12FS12		4.476	.646	87.882
FG>12FS>12		8.056	.778	3152.654
FG12012	.075	.501		1.078
FG120>12		.704	.488	2.022
FG>12012		-.285	.876	.752
FG>120>12		1.562	.836	4.768
FS12012	.645	.514		1.906
FS120>12		.917	.504	2.502
FS>12012		1.325	.858	3.762
FS>120>12		1.606	.831	4.983

FG12MG	-.197	.462	.821
FG>12MG	-2.292	1.469	.101
FS12MS	-2.156	1.233	.116
FS>12MS	-2.672	.951	.069
O12MG	-.924	.326	.397
O>12MG	-1.620	.281	.198
O12MS	-.883	.351	.414
O>12MS	-.865	.378	.421

Table 10. Hypothetical Data from Table 3 Collapsed Over Z

(a) Observed Table

Y = 0 Y = 1 Y = Missing

X = 0 960 480 360

X = 1 480 960 360

(b) Expanded Table

M = 0 M = 1

Y = 0 Y = 1 Y = 0 Y = 1 Total

X = 0 960 480 ? ? 360

X = 1 480 960 ? ? 360

1. Fuchs (1982), Little and Rubin (1987), Baker and Laird (1988), Winship and Mare (1989), and Park and Brown (1994) provide a technical discussion of how missing data can be incorporated into loglinear models. In this paper we discuss the key ideas needed for the researcher to incorporate missing data in a loglinear analysis.
2. These programs can estimate loglinear models with arbitrary patterns of missing data. As a result, we do not need to consider the issues found in earlier literature as to whether the missing data pattern is monotone and/or whether the likelihood can be factored (Little and Rubin 1987).
3. Winship and Mare (1989) provide a more extensive analysis of this example. In particular, we show, that the estimated effect for the association between X and Z is biased when we add a missing data category to Table 3.
4. Econometric models for sample selection bias are examples of NINR models (Winship and Mare 1992).
5. A related set of models explicitly distinguishes between the

"true" father's schooling and the fallible reports of father's schooling provided by each sibling. This distinction can be incorporated into latent class models. We do not consider these models in this paper.