ITEM NONRESPONSE IN ATTITUDE SCALES: A LATENT VARIABLE APPROACH

Colm O'Muircheartaigh, Irini Moustaki, London School of Economics
Colm O'Muircheartaigh, Dept of Statistics, LSE, Houghton Street, London WC2A 2AE

Key Words: Nonresponse, latent trait model, mixed manifest variables, pattern models

1. Introduction

This paper deals with the problem of treating item nonresponse in the analysis of attitude scales. It uses the maximum likelihood estimation of a latent dimension for metric variables (Lawley and Maxwell 1971), and for binary variables with a latent trait model (Bartholomew 1987). Albanese and Knott (1992) extended the latent variable model for binary items to allow for item nonresponse for sets of binary items. Using the latent trait model for mixed (binary and metric) variables (Moustaki 1996) we develop a general model for item nonresponse.

Opinion researchers who depend on a single attitudinal question are taking considerable risks because of the unreliability of responses to any given item. It is more common in measuring attitude dimensions to administer to the respondent (or to have the respondent complete) a battery of items. The battery will typically comprise a set of questions each of which attempts to tap a particular aspect of the attitude dimension under investigation. The responses to the set of questions are then combined in some way to provide an indication of the individual's position on the dimension. The items in the battery may be binary, polytomous, ordinal, or metric. In general we wish to combine the scores on the items in order to give a single overall score for each individual.

Among the best known methods for identifying the dimensionality of a multivariate metric attitudinal data set is factor analysis. For metric variables, a maximum likelihood solution was put forward by Lawley and Maxwell in 1971. Bartholomew (1987) proposed a general approach for fitting latent variable models on metric or categorical/ binary items. The score for each individual on the scale (a score corresponding to each possible pattern of responses) is obtained by either calculating the posterior mean of the distribution of the latent trait, given the pattern of responses for the individual concerned or by calculating component scores which are weighted sum of the responses.

No survey ever attains 100% response. The most serious case is unit nonresponse, where the individual does not respond to any of the items in the battery. Less disabling, but nevertheless consequential, is item nonresponse, where the individual fails to respond on some, but not all, of the items. In this paper we present a methodology for dealing with item nonresponse for metric, binary, and mixed (metric and binary) scales.

The analysis we present may be viewed in the general framework of maximum likelihood models for nonresponse. These may usefully be divided into selection models and pattern mixture models (Little, 1993, 1994). The information about the pattern of missing data is incorporated in a stochastic matrix, M, which has the same dimension as the data matrix, X, and takes the value 0 when an item is missing and a 1 when there is a response for the item; this matrix is called the missing-data indicator matrix. Both methods define the joint distribution of the complete data and the missing data mechanism. Different factorization of this joint distribution lead to the two different approaches. In one case, the factorization is of the complete data model and the missing-data mechanism model; this gives the selection model approach. In the second case the data are conditioned on the missing data mechanism; these are pattern mixture models. If the missing data mechanism does not depend on X, or if the missingness depends on the observed data but not on the unobserved values of X, then the data are missing completely at random (MCAR) or at random (MAR) and the nonresponse can be said to be ignorable. In non-technical terms this means that it is possible in principle to recover from the observed data enough information to reconstruct the unobserved data. For non-ignorable models, on the other hand, where the missing (unobserved) data depend on something other than the observed data, some additional constraints or assumptions are necessary to permit the missing data to be incorporated into the analysis.

In the case of the selection models this is brought about by specifying a form for the distribution of the missing data mechanism. For pattern mixture models, the likelihood is a function of the observed values only and hence parameters assigned to the missing data cannot be estimated directly. Restrictions are imposed for the estimation of these parameters.

There are two special features of our analysis. First, most approaches to nonresponse treat a single variable as the dependent variable in the analysis and the focus is on estimating missing values for this variable; in this sense the treatment is asymmetric. In the particular case of attitude measurement or scale construction, however, the status of each of the variables in the system is interchangeable; the analysis is in this sense symmetric. Thus as long as a case (an individual) has data for at least one of the variables in the system, that individual is a case of item nonresponse and not unit nonresponse.

The other special feature of our problem is that there is an underlying model for the relationship among the variables being measured. We assume that there is a single attitude dimension z_a (later work will extend the treatment to apply to multidimensional data) underlying all the items. We assume that there is a second dimension, response propensity, z_r on which individuals in the population vary. This variation may be due partly to differences in the basic response propensity of the individual; it may arise from differences from one item to another in the scale; and it may be influenced by the individual's position on the attitude continuum. These assumptions produce a non-ignorable nonresponse model and are consonant with how a survey researcher would think about nonresponse in this context. Using these two features we propose an approach which can be described as a *pattern model* approach.

We use the nature of the latent variable modelling structure to permit estimation of the parameters of the model. The probability of giving a response may depend on the individual's position both on z_a and z_r; the response itself depends only on the individual's position on z_a. Thus, conditional on the individual's having responded, the response itself depends only on the individual's position on z_a. Put another way, given the individual's position on the attitude dimension, his/her response to an item does not depend on his/her propensity to respond. In order to combine these two dimensions in the analysis it is necessary to be able to deal with a mixture of variable types in the same model (a mixed model). The attitude items may be metric, but the response items (response, nonresponse) will always be binary (we intend to extend the classification of nonresponse to the polytomous case in a later paper). The extended model was originally developed to deal with attitude scales in which the items themselves are a mixture of binary and metric items.

2. Modelling nonresponse

Albanese and Knott (1992) proposed a model for handling missing values in the analysis of attitudinal binary items with a latent trait model. This model does not distinguish between different sources of item nonresponse. Their model has been developed for binary manifest variables. This model is extended here for the case where we have either only metric manifest variables or mixed manifest variables; in both cases a mixed model is required to handle the nonresponse. Suppose that we have r manifest items to analyze and there is a proportion of nonresponse in each item. We create r pseudo items as follows, when an individual gives a response then the pseudo item for this individual will take the value one, when an individual do not respond to this item then the pseudo item will take the value zero. We now fit a two factor model on the $(2\times r)$ items. In a sense the first r items (the attitudinal items) provide us with information about attitude and the next r items (the response/nonresponse items) provide us information about propensity to express an opinion and they are called response propensity items. We proceed by fitting a two factors latent trait model on the $2 \times r$ items.
Equivalent to the results presented in Albanese and Knott (1992) we can break the response function into two layers.

For each (metric) attitude item:

$$(w_i \mid z_a, z_r, w_i \neq 9) \sim N(\mu_i+\lambda_{i1}z_a, \Psi_{ii}), \; i = 1,\cdots,r \tag{1}$$

For each response (pseudo) item:

$$Pr(w_i \neq 9 \mid z_a, z_r) = \pi_{ri}(z_a, z_r) \quad i = 1,\cdots,r \tag{2}$$

and

$$Pr(w_i = 9 \mid z_a, z_r) = 1 - \pi_{ri}(z_a, z_r) \tag{3}$$

Where for our model:

$$\text{logit}\pi_{ri}(z_a, z_r) = e_{i0} + e_{i1}z_a + e_{i2}z_r$$

The coefficient e_{i1} shows how the log of the odds of response increases or decreases with respect to the position of the individual on the attitude dimension. For the case of mixed (metric and binary) manifest variables additional equations are needed for the binary items. The log-likelihoods for estimation are maximized using the EM algorithm described in Moustaki (1996). The model can be fitted with the program LATENT (Moustaki 1995).

Interpretation of the model

Our primary interest is in finding a way to obtain information on how nonresponse represents attitude; for this we consider the posterior distribution of the attitude latent variable given each item. From the model parameters we obtain information on how attitude is related to propensity to respond and also information on how likely or unlikely we are to get a response for an item. But we are also interested in obtaining information about the missing values and what they represent in our sample. We propose here to look for each item at the posterior distribution of the attitude latent variable z_a given the possible responses for that item. So for binary items we are interested in observing the relative position of the $h(z_a \mid v_i = 9)$ with respect to $h(z_a \mid v_i = 0)$ and $h(z_a \mid v_i = 1)$ and for metric variables we look at the relative position of $h(z_a \mid w_i = 9)$ with respect to the three quartiles.

3. Applications

In this section we will present the results we found when we fitted the models on a number of data sets with missing values. We look first at some artificial examples in order to illustrate how the model works in cases where we know the structure of the data. The posterior analysis presented above will be looked at for all the data sets. We are first interested in using the model to score the missing value for an item relative to the other responses for that item and secondly to use this information to rank individuals on the attitude latent dimension.

3.1. Guttman scales

In fitting the models to standard Guttman scales (artificial data), we found that the model correctly identified item nonresponses that fell in unambiguous points on the scale. Where the missing value could have arisen from two different patterns of responses, the model placed the nonresponse between the relevant values. We illustrate this situation below in a data set containing a number of items that form a Guttman scale that includes a non-scale type. The data set used is given in Table 1. The pattern containing missing values [1 1 0 9] could arise from either response pattern 3 [1 1 0 1] or response pattern 4 [1 1 0 0]. This will create an ambiguity in the likely meaning of the nonresponses.

Table 1: Non-scale type 1

Response pattern	frequency
1 1 1 1	70
1 1 1 0	70
1 1 0 1	70
1 1 0 0	70
1 1 0 9	70
1 0 0 0	70
0 0 0 0	70

Looking at figure 1 we see that the $h(z_a \mid v_4 = 9)$ is somewhere in the middle between 0 and 1. Thus the model reflects the ambiguity in the data appropriately. When, however, we reinforced the data set with a metric variable (range 1-5) strongly correlated with the binary items the ambiguity in the meaning of the missing values disappears. The data set is given in Table 2. The model correctly assessed this new data set (see figure 2), and estimated the missing values in this case to be close to 0. We considered the discrepancies between the observed and expected two- and three-way margins of the attitudinal and the expression (pseudo) items. These discrepancies are measured with the statistic given by $(O-E)^2/E$. The discrepancies on the expression and the attitude items are very small.

Table 2: Non-scale type 2

Response pattern	frequency
1 1 1 1 5	70
1 1 1 0 4	70
1 1 0 1 4	70
1 1 0 0 3	70
1 1 0 9 3	70
1 0 0 0 2	70
0 0 0 0 1	70

3.2. Sexual attitudes: BSA 1990

The data set used here have been extracted from the British Social Attitudes, 1990, Survey. There were 1270 individuals who were asked questions on sexual relationships.

1...Now I would like you to tell me whether, in your opinion, it is acceptable for a homosexual person to be a teacher at a school? [SchTGay]

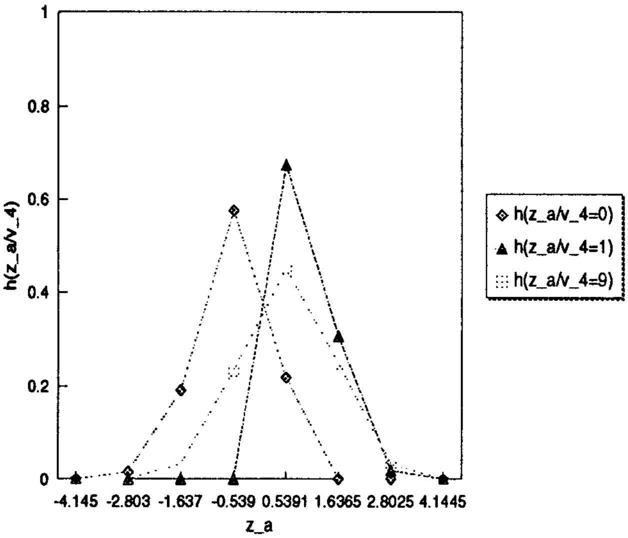

Figure 1: Non-scale type 1, posterior probabilities

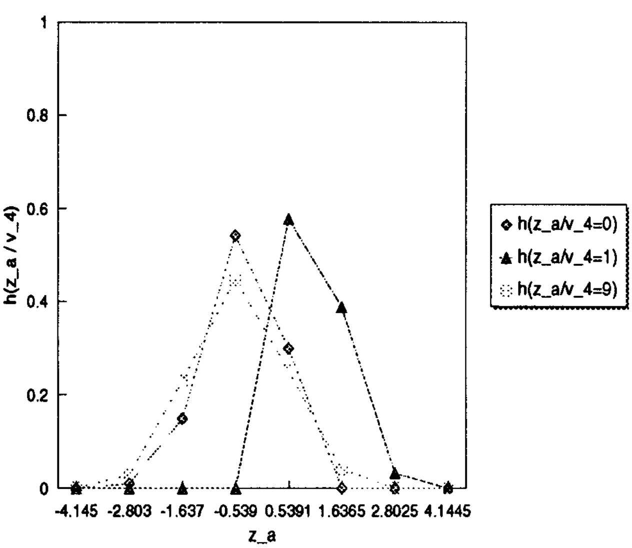

Figure 2: Non-scale type 2, posterior probabilities

2...Now I would like you to tell me whether, in your opinion, it is acceptable for a homosexual person to be a teacher in a college or a university? [HEdTGay]
3...Now I would like you to tell me whether, in your opinion, it is acceptable for a homosexual person to hold a responsible position in public life? [PLifeGay]
4...What about sexual relations between two adults of the same sex? [HomSex]
The items 1 to 3 are binary items with response categories 1 for agree and 0 for disagree and item 4 is a five point scale item with responses "always wrong", "mostly wrong", "sometimes wrong", "rarely wrong" and "not wrong at all", treated here as discrete. The percentage of nonresponse for item 1 is 1.8% for item 2 is 1.9% for item 3 is 2% and for item 4 is 0.8%.

From the tables with the one- two- and three-way margins the fit of the model looks satisfactory. The formulation of the model allows attitude to affect expression; a measure of this effect can be obtained by looking at the coefficients e_{i1}. The values of these coefficients should be considered in the context of the posterior probabilities for the four items given below. Item 1 has a value of $e_{i1} = -0.08$ that indicates that attitude is not related to propensity to respond for this item and as a result no information can be obtained for attitude from nonresponse. From the posterior analysis discussed above we find that the $h(z_a \mid v_i = 9)$ for item 1 is in the middle between 0 and 1 (see figure 3). Item 2 has a value of $e_{i1} = 0.94$ that indicates that the more positive attitude an individual has towards homosexuality the more likely s/he is to respond; thus an individual who fails to respond is more likely to be on the left side of the attitude scale, a score near 0 (see figure 4). A nonresponse to item 3 is even closer to 0 since the value of $e_{i1} = 1.20$ (see figure 5). For item 4, however, the value of $e_{i1} = -1.04$; this indicates that the more positive the attitude an individual has towards homosexuality the less likely s/he is to respond; thus an individual who fails to respond is more likely to be on the right side of the attitude scale, see figure 6. As this is a metric item with a range of 1-5, this means that such an individual is likely to have a score near 5. The scores of individuals on the attitude scale based on their whole response pattern is given in Table 3. A simple example may illustrate the use of the table. If we compare individuals who respond positively to the first three items (patterns beginning 1 1 1) we see that someone who has not responded to the fourth item scores almost the same as someone who gives the maximum response (5) to that item, and higher than those who give responses of 1, 2, 3, and 4.

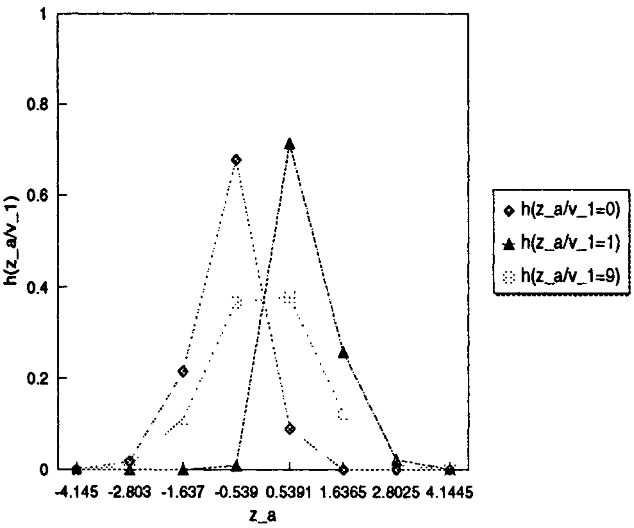

Figure 3: Item SchTGay, posterior probabilities

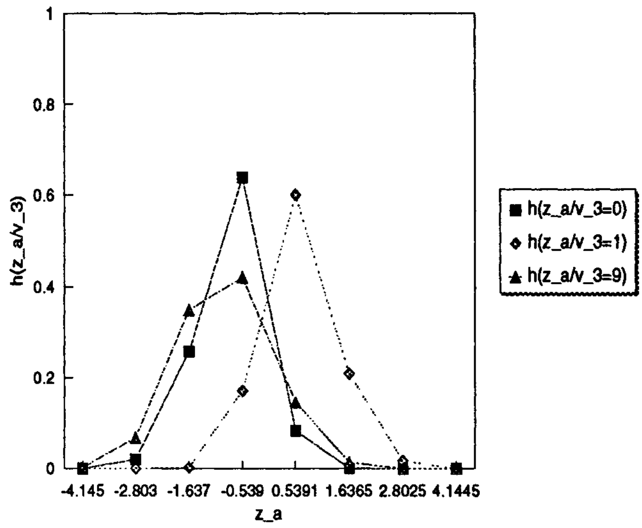

Figure 5: Item PLifeGay, posterior probabilities

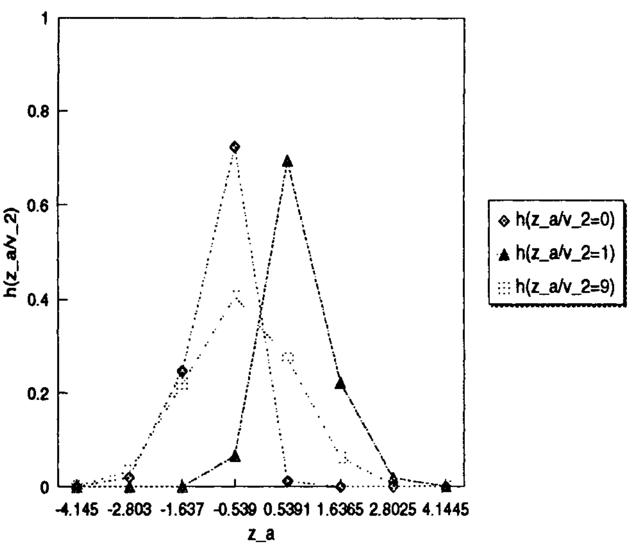

Figure 4: Item HEdTGay, posterior probabilities

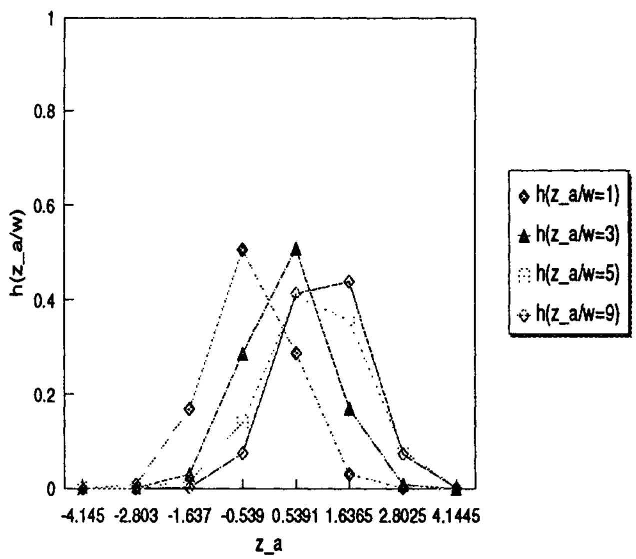

Figure 6: Item HomSex, posterior probabilities

Table 3: Posterior mean, sexual attitudes

$E(z_a \mid \mathbf{w}, \mathbf{v})$	responses	$E(z_a \mid \mathbf{w}, \mathbf{v})$	responses
-1.51	0991	0.00	0105
-1.20	0091	0.25	0111
-1.09	0901	0.26	1011
-0.92	0001	0.31	9191
-0.91	0902	0.33	9911
-0.81	0099	0.36	0112
-0.80	9001	0.36	1012
-0.78	0002	0.43	0113
-0.75	9991	0.47	0114
-0.68	0003	0.50	1015
-0.66	9901	0.53	1101
-0.62	0004	0.54	1102
-0.59	0005	0.55	1013
-0.55	0011	0.57	1104
-0.54	0012	0.57	1911
-0.53	0013	0.59	1105
-0.52	0014	0.62	1912
-0.50	0015	0.63	9111
-0.47	0911	0.67	1111
-0.46	9992	0.68	9913
-0.44	0101	0.76	1112
-0.37	0102	0.83	1195
-0.35	9903	0.88	9113
-0.34	0191	0.90	1113
-0.28	0103	1.08	9919
-0.23	0192	1.08	1114
-0.17	9909	1.32	1115
0.00	9101	1.35	1119

4. Comparison of the LVA with PMM

In this section we discuss possible similarities and dissimilarities between the formulation proposed here for the case of missing values in attitude scales using a latent variable approach and the formulation provided by pattern mixture models (PMM). The observed variable X_i takes the observed values or 9 (a nonresponse). There are p variables to be analyzed. We create a number of p pseudo items that take the value zero for the response 9 and 1 otherwise. The matrix with the pseudo items is the same as the missing-data indicator matrix M. Albanese and Knott (1992) model specification does not require specification of the joint distribution of the complete data X and the stochastic matrix M but it breaks down the response function into two layers. One layer defines the probability distribution of response to an item for an individual with position on the attitude latent variable z_a who responded to that item, an the other layer defines the probability that an individual responds to an item given his position on the two latent variables z_a and z_r.

The likelihood to be maximized is based on the joint distribution of the observed variables X where each variable takes either the response value, or 9 if there is not a response. The parameters in the model are all estimated using an E-M algorithm. In the examples presented by Little 1993, 1994 the model specification for the pattern mixture model requires the form of the distribution of X_{obs} given the missing values. The likelihood depends on the observed part of the data and so all the parameters depending on missing data are undefined. These problems are handled by imposing constraints on the undefined model parameters.

The latent variables fitted in the model provide us with indirect information about attitude from nonresponse as was shown in the section above. By introducing the latent variables we allow the missing data mechanism to depend on the observed values or the missing ones through the latent variables z_a and z_r respectively. The latent variable approach has the flexibility of incorporating into the analysis both the observed and the unobserved part of the data and allowing an estimation of the model parameters using the complete pattern of responses and nonresponses in the data. This *pattern model* approach does not require any further restrictive assumptions to permit estimation of the parameters.

References

Albanese, M. T. and M. Knott (1992). TWOMISS: a computer program for fitting a one- or two-factor logit-probit latent variable model to binary data when observations may be missing. Technical report, Statistics Department, London School of Economics and Political Science, England & Universidade Federal do Rio Grande do Sul, Brazil. The software will be available over EMAIL. Send inquiries to the authors. (EMAIL to M.KNOTT@LSE.AC.UK).

Bartholomew, D. J. (1987). *Latent Variable Models and Factor Analysis.* London: Charles Griffin & Co. Ltd.

Lawley, D. N. and A. E. Maxwell (1971). *Factor Analysis as a Statistical Method.* London: Butterworth.

Little, R. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association 88*, 125–134.

Little, R. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika 81*, 471–483.

Moustaki, I. (1995). LATENT: A computer program for fitting a one- or two- factor latent variable model to mixed observed items. Technical report, Statistics Department, London School of Economics and Political Science.

Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British Journal of Mathematical and Statistical Psychology 49.* forthcoming.