Mai Sherif Hafez, Irini Moustaki and Jouni Kuha

Analysis of multivariate longitudinal data subject to nonrandom dropout

Article (Accepted version) (Refereed)

Original citation:
Hafez, Mai Sherif, Moustaki, Irini and Kuha, Jouni (2015) *Analysis of multivariate longitudinal data subject to nonrandom dropout.* Structural Equation Modeling: A Multidisciplinary Journal, 22 (2). pp. 193-201. ISSN 1070-5511

DOI: 10.1080/10705511.2014.936086

Analysis of multivariate longitudinal data subject to nonrandom dropout

Mai Sherif Hafez, Irini Moustaki and Jouni Kuha
London School of Economics and Political Science[1]

Suggested running head: Nonrandom dropout

To appear in Structural Equation modeling

[1] Department of Statistics, Houghton Street, London WC2A 2AE, U.K. Email: m.m.hafez@lse.ac.uk, i.moustaki@lse.ac.uk (correspondence author), j.kuha@lse.ac.uk

Abstract

Longitudinal data are collected for studying changes across time. We consider multivariate longitudinal data where multiple observed variables, measured at each time point, are used as indicators for theoretical constructs (latent variables) of interest. A common problem in longitudinal studies is dropout, where subjects exit the study prematurely. Ignoring the dropout mechanism can lead to biased estimates, especially when the dropout is *nonrandom*. Our proposed approach uses latent variable models to capture the evolution of the latent phenomenon over time while also accounting for possibly nonrandom dropout. The dropout mechanism is modeled with a hazard function that depends on the latent variables and observed covariates. Different relationships among these variables and the dropout mechanism are studied via two model specifications. The proposed models are used to study people's perceptions on women's work using three questions from five waves from the British Household Panel Survey.

Keywords: structural equation modeling, ordinal variables,nonignorable dropout, weighted least squares, response propensity

1. Introduction

In this article, we consider latent variable modeling of multivariate longitudinal data subject to nonrandom dropout. Longitudinal data are collected for studying changes across time. Most of the existing research on longitudinal data focuses on repeated measures for one variable over time. Good starting points to the extensive literature on such univariate longitudinal data analysis are Diggle, Heagerty, Liang, and Zeger (2013), who give a thorough overview of different methods, and Verbeke and Molenberghs (2000), who provide a comprehensive treatment of linear mixed models for continuous longitudinal data.

However, in social science applications, including educational testing and psychometrics, the main interest is often in theoretical constructs, such as attitudes, behaviour or abilities, which cannot be directly measured. In that case, multiple observed variables ('items'), for example survey questions or items in an ability test, are used as indicators for the constructs, which are themselves treated as unobservable (latent) variables. The observed items and the latent variables are linked together by statistical latent variable models (see e.g. Skrondal and Rabe-Hesketh (2004) and Bartholomew, Knott, and Moustaki (2011) for overviews). In particular, in this paper we consider models which treat the items as ordinal, because such variables are often met in social surveys.

When the interest lies in how the latent constructs change across time, the same items are measured at different time points, thus resulting in multivariate longitudinal data. Models for such data have been proposed by, for example, Fieuws and Verbeke (2004, 2006), Dunson (2003), and Cagnone, Moustaki, and Vasdekis (2009), who model the associations of the latent and observed variables across time using random effects and/or latent variables.

A common problem in longitudinal studies is dropout, where subjects exit the study prematurely. A crucial question for the analysis is whether or not those who drop out are systematically different from the ones who remain till the end of the study. In the widely used terminology due to Rubin (1976), data are considered missing completely at random (MCAR) if the missingness (in our case dropout) is independent of both observed and unobserved data, missing at random (MAR) if the missingness depends on the observed data but not on the unobserved, and missing not at random (MNAR) if it depends on unobserved data. When modeling longitudinal data, the joint density function of both the measurement and dropout processes is considered. If the dropout is at random (i.e. MAR), and the parameters of the dropout process are distinct from those of the models for the latent variables and their measurements (an assumption we make throughout), the dropout is said to be ignorable and a valid analysis can be based on a likelihood that ignores the dropout mechanism. However, it is not always easy to justify the assumption of random dropout. If it does not hold, the dropout mechanism should be incorporated in the analysis of the data, as ignoring it may lead to biased estimates of the parameters of interest.

There are three general approaches for modeling univariate longitudinal data subject to dropout, the first two of which are most common. *Selection models* factorise the joint density into the product of the marginal density of the measurement process and the conditional density of the missingness mechanism given the measurement. In their key paper on selection models for non-ignorable dropout, Diggle and Kenward (1994) combine a multivariate Gaussian linear model for the measurement process with a logistic dropout model. Molenberghs, Kenward, and Lesaffre (1997) use a similar framework to model dropout probabilities when the variable of interest is ordinal. Jansen, Beunckens, Molenberghs, Verbeke, and Mallinckrodt (2006) also study non-Gaussian outcomes such as binary, categorical or count data. They consider both generalized linear mixed models, for which the parameters can be estimated using maximum likelihood, and marginal models estimated through generalized estimating equations, which is a nonlikelihood method and hence requires a modification to be valid under MAR.

Pattern-mixture models (Little, 1993) are an alternative to selection models. They factorise the joint density in the opposite way, that is as the product of the marginal density of the dropout mechanism, and the conditional density of the measurement process given the dropout. In other words, the measurement process is defined over different dropout patterns.

The third general approach for modeling dropout are *shared-parameter models*, in which both the measurement process and dropout are influenced by a latent variable or random effect (e.g. Wu & Carroll, 1988; Wu & Bailey, 1989; Henderson, Diggle, & Dobson, 2000). A shared parameter model is thus a selection model which is also conditional on a latent variable. This specification allows the dropout to be non-ignorable given the observed data only, but ignorable given also the latent variables. Roy (2003) introduced a shared-parameter model in which the dependence between the measurement process and time of dropout is due to a shared latent variable that is assumed to be discrete, so that the marginal distribution of the measurement is a mixture over the dropout classes of the latent variable. Dantan, Proust-Lima, Letenneur, and Jacqmin-Gadda (2008) compare pattern-mixture models and latent class models in dealing with informative dropout.

Our approach to handling dropout in multivariate longitudinal data draws on ideas of shared parameter models for univariate longitudinal data, and on previous work on modeling non-ignorable item nonresponse in multivariate cross-sectional data. Early examples of the latter are Knott, Albanese, and Galbraith (1990) and O'Muircheartaigh and Moustaki (1999), who present a latent variable approach that allows missing values to be included in the analysis and information about latent attitudes to be inferred from nonresponse. They propose two latent dimensions, one to summarise the attitude and the other to summarise response propensity. For each observed variable, an indicator variable for responding is created, taking the value 1 if the individual responds and 0 if he or she does not respond. The attitude items are explained by the attitudinal latent variable, and the binary response items depend both on the attitudinal variable and the response propensity latent variable, thus allowing for non-ignorable missingness. Holman and Glas (2005) use reformulations of the models of O'Muircheartaigh and Moustaki (1999) to assess the extent to which the missing data are non-ignorable. Within the same framework, Moustaki and Knott (2000) present a latent variable model for binary and nominal observed items which includes covariate effects on attitudinal and response propensity items. In our study, we extend this approach to the longitudinal case with nonrandom dropout.

The models developed in this paper are latent variable models in which a continuous latent variable is used at each time point to explain the associations among multiple observed response items. Random effects are included to account for repetition of items over time. For modeling dropout, we introduce dropout indicators which are modeled with a hazard function. Different structures among the latent variables and the dropout mechanism are explored in two different model specifications which allow attitudes and covariates to affect both the latent variable and the dropout indicators.

We apply the proposed models to study the evolution of people's attitudes towards women's work, using data from the British Household Panel Survey. Five waves of the survey (1993, 95, 97, 99, 2001) are considered here. Dropout occurs in all waves but the first one. We analyse three survey items, which are worded as follows: 'A woman and her family would all be happier if she goes out to work' (labelled 'Family' below), 'Both the husband and wife should contribute to the household income' ('Contribution'), and 'Having a full-time job is the best way for a woman to be an independent person' ('Independent'). For each of them, the response options are 'Strongly agree', 'Agree', 'Neither agree nor disagree', 'Disagree', and 'Strongly disagree'. The attitudinal latent variable will be defined so that the higher an individual scores on the latent variable, the more conservative are his or her views towards women's work. The analysis aims to explore how much each of the three items contributes to measuring this attitude and how the attitude evolves over the nine-year period, accounting for dropout by incorporating the dropout mechanism in the model.

Section 2 lays out the general framework for the proposed model and presents two possible model specifications, and Section 3 describes the results from the data analysis. Final comments and conclusions are given in Section 4.

2. A latent variable model for multivariate longitudinal data subject to dropout

A latent variable model is first specified for the complete-case multivariate data, disregarding dropout. This model is formed of two parts: the measurement part in which the observed variables are explained by a latent variable at each time point, and the structural part which defines relationships among the latent variables over time. Having specified this model for the complete data, we then define models for the dropout mechanism with a hazard function. Finally, the link between attitudes and dropout is specified.

2.1 Modeling the observed indicators: The measurement model

We will consider ordinal items as they are among the most common type of items used for measuring attitudes in social surveys. Suppressing the index for a subject (e.g. survey respondent) for convenience, let $\mathbf{y}_t = (y_{1t}, y_{2t}, ..., y_{pt})$ be $p \times 1$ vectors of observed ordinal variables for a single subject at times $t = 1, 2, ..., T$. Let c_{it} denote the number of categories for y_{it}, the ith variable $(i = 1, 2, ..., p)$, at time t. It is assumed that each y_{it} is a manifestation of an underlying unobserved continuous variable y^*_{it}. For an ordinal variable y_{it} with c_{it} categories, its relationship with y^*_{it} is given in Jöreskog (2005) as

$$y_{it} = s \Leftrightarrow \tau^{(i)}_{s-1} < y^*_{it} \leq \tau^{(i)}_{s}, \quad s = 1, \cdots, c_{it}, \tag{1}$$

where $\tau^{(i)}_0 = -\infty$, $\tau^{(i)}_1 < \tau^{(i)}_2 < \ldots < \tau^{(i)}_{c_{it-1}}$, and $\tau^{(i)}_{c_{it}} = \infty$ are known as thresholds. There are $c_{it} - 1$ estimable thresholds for an ordinal variable with c_{it} categories. The underlying variable y^*_{it} is assumed to have a standard normal distribution.

The items $\mathbf{y}_t = (y_{1t}, y_{2t}, ..., y_{pt})$ at each time t are regarded as measures of a continuous attitudinal time-dependent latent variable z_{a_t}, which is assumed to be normally distributed. For simplicity, the model below is presented assuming that the items are unidimensional (i.e. one latent variable is sufficient to explain dependencies among items at a given time point), but it can be extended to accommodate more latent variables. The measurement model for z_{a_t} at each time t is the classical factor analysis model

$$y^*_{it} = \lambda_i z_{a_t} + u_i + \varepsilon_{it}; \qquad i = 1, ..., p\,;\ t = 1, ..., T, \tag{2}$$

where λ_i is the loading of the latent variable z_{a_t} on y^*_{it}, u_i is an item-specific random effect, and ε_{it} is a random error. In this model, associations among different items at the same time (y^*_{it}, y^*_{jt} for $i \neq j$) are explained by the dependence on the common latent variable z_{a_t}, while associations between the values of the same item measured at different time points ($y^*_{it}, y^*_{it'}$ for $t \neq t'$) are explained both by the covariance between corresponding attitudinal latent variables $(z_{a_t}, z_{a_{t'}})$ and the item-specific random effect u_i (equivalently, errors ε_{it} of the same item across time could be allowed to correlate rather than introducing the random effects). It is assumed that the random effects u_i are independently normally distributed as $u_i \sim N(0, \sigma^2_{u_i})$ for $i = 1, ..., p$, and that ε_{it} are independent and normally distributed as $\varepsilon_{it} \sim N(0, \upsilon^2_{\varepsilon it})$ for $i = 1, ..., p$ and $t = 1, ..., T$, where $\upsilon^2_{\varepsilon it} = 1 - (\lambda^2_i \operatorname{var}(z_{a_t}) + \sigma^2_{u_i})$ since each y^*_{it} is assumed to have a standard normal distribution. The error terms ε_{it} and random effects u_i are assumed to be uncorrelated.

In the measurement model (1)–(2) we have imposed the assumption of *invariance of measurement* across time for each item $i = 1, \ldots, p$, by constraining the thresholds $\tau^{(i)}_s$ (for each $s = 1, \ldots, c_{it}$) and the loading λ_i for each $i = 1, \ldots, p$ to be the same at all time points $t = 1, \ldots, T$. The advantages of this constraint are both technical and conceptual. On the technical side, it yields a more parsimonious model and avoids some possible identification problems that may arise with increasing the number of time points (Bijleveld, Mooijaart, van der Kamp, & van der Kloot, 1998). The conceptual advantage is clearer interpretation of the model results. If the loadings and thresholds are not constrained to be time-invariant, we cannot guarantee that the latent variable has the same interpretation at each time point.

In order to set the scale for the time-dependent attitude latent variables, and for their variances $\sigma^2_1, ..., \sigma^2_T$ to be estimable, the loading λ_1 on the first observed variable y_{1t} is set to 1. Also, the loadings of each random effect u_i on an item at different occasions $y_{i1}, ..., y_{iT}$ are all set to 1, thus making all occasions contribute equally to the random effect. Then the variances $\sigma^2_{u_i}$ of the random effects are left to be estimated.

This model specification has been introduced by Dunson (2003) in a generalized linear latent variable model framework for different response types where Markov Chain Monte Carlo (MCMC) methods were used for estimation. Cagnone et al. (2009) propose a full-information maximum likelihood estimation method for the same model specification with ordinal variables. Cai (2010) develops an EM algorithm for full-information maximum marginal likelihood estimation that is computationally efficient due to the use of a dimension reduction technique of the latent variable space for the two-tier item factor analysis model, which fits into this model specification. Composite likelihood approaches have also been proposed to reduce estimation complexity for this type of models (see Vasdekis, Cagnone, & Moustaki, 2012). We consider estimation using diagonally weighted least squares (DWLS), and weighted least squares (WLS) for the same model specification for multivariate longitudinal data within a structural equation modeling (SEM) framework where ordinal variables are treated using underlying continuous variables.

2.2 Modeling the latent variables: The structural model

The structural part of the model addresses the question: how should the attitudinal latent variables be linked in order to capture the longitudinal nature of the data? Throughout, we will assume that the possible measurement occasions $t = 1, \ldots, T$ are the same for every subject, and evenly spaced in time. We then specify that the $T \times 1$ vector of attitude latent variables $\mathbf{z}_a = (z_{a_1}, ..., z_{a_T})'$ follows a multivariate normal distribution $\mathbf{z}_a \sim MVN_{(T)}(\boldsymbol{\mu}, \boldsymbol{\Gamma})$ where $\boldsymbol{\mu}$ is a vector of means and $\boldsymbol{\Gamma}$ a covariance matrix with diagonal elements σ_t^2 representing the variances of the latent variables, and off-diagonal elements $\sigma_{tt'}$ their covariances such that $\sigma_{tt'}$ is the covariance between z_{a_t} and $z_{a_{t'}}$. The values of these parameters may be unconstrained, or depend further on the model specification, as defined below. For example, it is logical to expect that attitudes are more strongly correlated when they are measured at closer time points, in which case $\sigma_{tt'}$ should be higher when t and t' are close to each other. For identification, the mean of z_{a_1} is set to 0.

A specification for the structural part which takes the time ordering explicitly into account is the first-order autoregressive [AR(1)] structure where $z_{a_1} \sim N(0, \sigma_1^2)$ and

$$z_{a_t} = \alpha_t + \phi z_{a_{t-1}} + \delta_t, \quad t = 2, ..., T, \tag{3}$$

where α_t is an intercept, ϕ a regression coefficient representing the dependence of the attitude at time t on that at the previous occasion $t-1$, and $\delta_t \sim N(0, \upsilon_{\delta t}^2)$ is a random error which is uncorrelated with $z_{a_1}, ..., z_{a_{t-1}}$. This formulation explicitly captures the time ordering in the data, by presenting the model as a sequence of conditional distributions rather than a joint distribution with a completely free correlation matrix $\boldsymbol{\Gamma}$. It expresses the dynamic nature of the latent attitude variable (Dunson, 2003; Cagnone et al., 2009) and accounts for the serial correlation in it in a form where the latent variable at time point 3, say, is only related to that measured at time 1 via the latent variable at time 2. Another alternative specification would be a random effects model in which a random intercept and possibly a random slope affect the time-dependent latent variables as in a standard growth mixture model for observed repeated measures; for example, see Muthén and Masyn (2005) and Muthén, Asparouhov, Hunter, and Leuchter (2011). However, this type of model is not considered here.

More generally, we may also be interested in studying the associations between the attitudinal latent variables and observed covariates (explanatory) variables, such as demographic and socioeconomic characteristics of survey respondents. Let $\mathbf{x}_t$ denote a vector of such covariates, noting that some components of $\mathbf{x}_t$ (e.g. sex and race) may be constant over time while others (e.g. marital status and health condition) may be time-varying. In this case, the AR(1) structure in (3) can be extended to include covariates, as

$$z_{a_t} = \alpha_t + \phi z_{a_{t-1}} + \boldsymbol{\theta}'\mathbf{x}_t + \delta_t, \quad t = 2, ..., T, \tag{4}$$

where $\boldsymbol{\theta}$ is a vector of regression coefficients for $\mathbf{x}_t$.

2.3 Modeling the dropout

Dropout is a form of missing data in which a respondent in a longitudinal study fails to respond at a given occasion and never comes back to the study. It contrasts with ‘intermittent’ missingness where an individual who does not show up at a given occasion may return at a subsequent one. Dropout is typically the most common form of missingness in longitudinal studies. We will focus solely on it, and assume that there is no intermittent missingness in the data. We also assume that, at each time point, variables for a respondent are either fully observed or totally missing, i.e. that there is no item nonresponse. We define the probability that a respondent drops out at time t, given that they have remained in the study up to and including time $t-1$, by the hazard function $h_t = P(K = t \mid K \geq t)$, $t = 2, ..., T$, where K is a discrete random variable that indicates the time of dropout. We also define a set of dropout indicators d_t, $t = 1, \ldots, T$, such that $d_t = 0$ when $\mathbf{y}_t$ is observed and $d_t = 1$ if a respondent drops out at time t (Muthén & Masyn, 2005). After the time of dropout, d_t itself is regarded as missing and can be set to an arbitrary value such as 999. We treat the observations at the first occasion as complete data, so that $d_1 = 0$ for all respondents, and define $\mathbf{d} = (d_2, ..., d_T)$. For an example with three waves ($T = 3$), an individual will have $\mathbf{d} = (0, 0)$ if they show up on all three occasions, $\mathbf{d} = (0,\ 1)$ if they drop out on the third occasion, and $\mathbf{d} = (1,\ 999)$ if they drop out on the second occasion. With this notation, the hazard function can also be expressed as

$$h_t = P(K = t \mid K \geq t) = P(d_t = 1), \quad t = 2, ..., T.$$

In the more general case of intermittent missingness we could define binary missingness indicators such that the indicator d_t at time t has the value 0 if $\mathbf{y}_t$ is observed and 1 if it is missing. In that case, the missingness indicators may be assumed to measure a single latent variable z_{d_t} which summarises an individual's 'response propensity'. Such a propensity may also be thought to exist in our case, where only dropout is considered. However, since the dropout indicators are created from a single variable (time of dropout), this latent propensity cannot be separately identified. Nevertheless, we will still employ such z_{d_t} as a convenient computational and presentational device, but with a formulation where they have a conditional variance of 0, given the attitude latent variables z_{a_t} and (possibly) covariates $\mathbf{x}_t$ (Muthén & Masyn, 2005). This means that z_{d_t} will be deterministic functions of z_{a_t} and $\mathbf{x}_t$, which will then affect the dropout indicators via z_{d_t}.

In the same way as for the observed items $\mathbf{y}_t$ in equation (1), we assume a set of continuous variables $\mathbf{d}^* = (d_2^*, \ldots, d_T^*)$ to underlie the set of dropout indicators $\mathbf{d} = (d_2, \ldots, d_T)$. Each of the d_t^* is assumed to have a standard normal distribution and to be modeled as

$$d_t^* = \lambda_{d_t} z_{d_t} + \varepsilon_{d_t}, \quad t = 2, ..., T, \tag{5}$$

where λ_{d_t} is the loading of z_{d_t} on the dropout variable at time t, and $\varepsilon_{d_t} \sim N(0, \sigma^2_{\varepsilon_{dt}})$ is a random error, with $\sigma^2_{\epsilon_{dt}} = 1 - \lambda^2_{d_t} \text{var}(z_{d_t})$. Since the missingness indicators are all binary, only one threshold τ_{dt} is estimated for each variable d_t^*.

We will consider two special cases of this model. In the first, we take $z_{d_t} = z_{a_{t-1}}$ for $t = 2, \ldots, T$. Model (5) then becomes

$$d_t^* = \lambda_{d_t} z_{a_{t-1}} + \varepsilon_{d_t}, \quad t = 2, ..., T. \tag{6}$$

In this formulation, the probability of dropping out at a given time point depends only on the value of the latent attitude variable at the immediately preceding time point. The dropout indicators are thus in effect treated just like further 'measures' of the attitude. Because the loadings λ_{d_t} can vary with t, the effect of attitude on dropout may depend on time.

In our second dropout model we define $z_{d_t} = z_d$ instead as a time-constant quantity which depends on the attitude only through its value z_{a_1} at the first time point. In this formulation we also allow for the possibility that the response propensity depends also on covariates $\mathbf{x}_1$ measured at the first time point. We thus define $z_d = \beta z_{a1} + \boldsymbol{\omega}'_d \mathbf{x}_1$, where β is a regression coefficient representing the dependence of the dropout 'latent variable' z_d on the attitude latent variable z_{a_1} at the first time point, and $\boldsymbol{\omega}_d$ is a vector of the regression coefficients of covariates $\mathbf{x}_1$ similarly. Furthermore, in (5) we take $\lambda_{d_t} = 1$ for all t, to obtain

$$d^*_t = \beta z_{a1} + \boldsymbol{\omega}'_d \mathbf{x}_1 + \varepsilon_{d_t}, \quad t = 2, ..., T. \tag{7}$$

Here the time-constant dropout variable z_d is regressed solely on z_{a_1} in order to avoid a multicollinearity problem that is very likely to occur if z_d was regressed on other attitude latent variables as well, due to the high correlation expected between the latent variable across different time points. Attitude at the first time is particularly chosen because it is the only occasion with complete data, and because it avoids a specification where dropout at time t would depend on attitude at future time points. Following the same argument, dropout is also regressed only on covariates measured at the first time.

The specification of the dropout models determines the nature and informativeness of the dropout. The missing data will be MCAR if the models for d^*_t depend neither on the latent variables z_{a_t} nor covariates $\mathbf{x}_t$, and MAR if they depend on $\mathbf{x}_t$ but not on z_{a_t}. In particular, any model where d^*_t depends directly on the latent attitudes z_{a_t} implies nonrandom dropout, i.e. that the data are missing not at random (MNAR) and the dropout process is thus non-ignorable. In model (6), non-ignorability holds unless λ_{d_t} are 0 for all $t = 2, \ldots, T$, and in model (7) it holds unless $\beta = 0$.

When dropout is non-ignorable, a model for it needs to be incorporated in the estimation in order to obtain valid estimates for the parameters of interest in the structural and measurement models. For multivariate longitudinal data, unlike in many other situations, this can in fact be done without further unverifiable assumptions. In other words, combining the elements described above it is possible to fit models which combine multivariate longitudinal models for the latent attitude variables of interest with models for non-ignorable dropout. In the next section we discuss such joint models in more detail.

2.4 Joint models for attitudes, measurements and dropout

Having set the general layout of the model, we now look into two particular specifications of it. In both of them, the measurement model of the observed items y_{it} is defined by equations (1) and (2), and the corresponding assumptions. Differences lie in the definitions of the structural and dropout parts of the model, and the relationship between them.

The first model specification allows for the simple choice of a free mean structure and correlation matrix for the attitudinal latent variables $\mathbf{z}_a = (z_{a_1}, ..., z_{a_T})'$ at different time points. In other words, we assume a multivariate normal distribution $\mathbf{z}_a \sim MVN_{(T)}(\boldsymbol{\mu}, \boldsymbol{\Gamma})$ with $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$ unconstrained. For incorporating dropout, we assume model (6) where the attitudinal latent variable $z_{a_{t-1}}$ at each previous time point is allowed to directly affect the dropout at the next one. The parameters of this dropout model are the thresholds τ_{dt} and loadings λ_{dt} for $t = 2, \ldots, T$, with non-zero λ_{dt} indicating nonrandom dropout. Figure (1) gives an illustration of this model by a path diagram for an example with three time points.

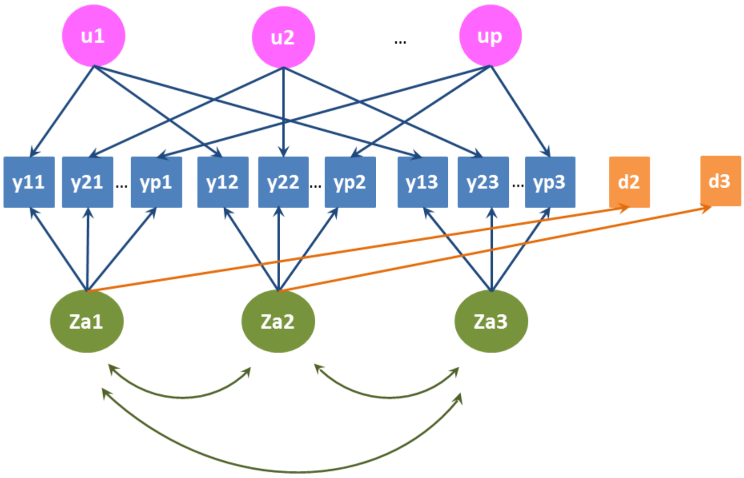

Figure 1. Path diagram for the first model specification ('Model 1').

The second model specification assumes a first-order autoregressive structure among the latent variables $\mathbf{z}_a$, as presented in equation (3), instead of freely correlating them. With the attitude at the first time point also assumed to be normally distributed as $z_{a_1} \sim N(0, \sigma_1^2)$, this model too implies that $\mathbf{z}_a$ follows a multivariate normal distribution, but now with the covariance matrix $\mathbf{\Gamma}$ being a constrained function of the parameters ϕ, σ_1^2 and $\upsilon_{\delta 2}^2, \ldots, \upsilon_{\delta T}^2$, and the mean vector unconstrained and depending on the parameters $\alpha_2, \ldots, \alpha_T$ and ϕ. For this model specification we also examine the extension of the structural model by including in it covariates $\mathbf{x}_t$ with coefficients $\boldsymbol{\theta}$, as shown in equation (4).

For the dropout model in the second model specification, we assume a model where the underlying dropout variables are modeled as a function of the dropout 'latent variable' z_d which in turn is determined by the attitude latent variable z_{a_1} and covariates $\mathbf{x}_1$ at the first time point, thus resulting in the dropout model (7). Figure (2) gives an illustration of the joint model for the second model specification, for an example with three time points.

In the second specification the parameters of the dropout model are the thresholds τ_{dt} $(t = 2, \ldots, T)$ and the regression coefficients β and $\boldsymbol{\omega}_d$, with non-zero β indicating nonrandom dropout. These parameters are to be estimated, along with the parameters of the measurement model (including the variances of the random effects u_i) and the structural model.

3. Data Analysis

The data used in this analysis come from five waves of the British Household Panel Survey (BHPS). We consider three survey questions as given in Section 1, which are treated as measures of a respondent's attitude towards women's work. The sample size of individuals who gave complete answers in the first wave considered here (year 1993) is 5819. In the second wave, with 10% dropout the sample size decreases to 5227, and in the third wave, a further 6% dropout reduces it to 4901. Dropout continues at each wave until the sample

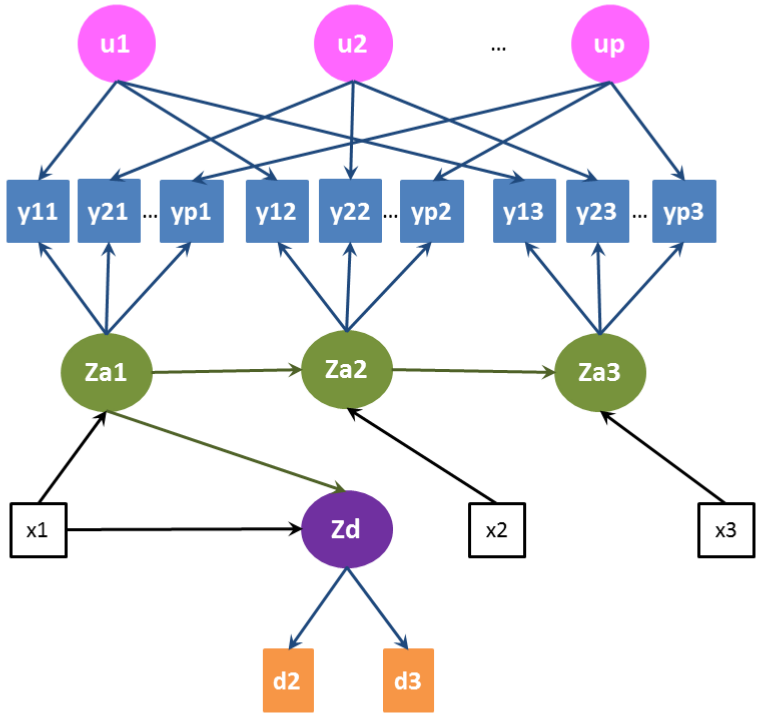

Figure 2. Path diagram for the second model specification ('Model 2') with covariates.

size becomes 4296 at the last wave considered here (year 2001), constituting approximately 74% of the original sample size. Results from the two different model specifications outlined previously are compared. Moreover, covariates are introduced and their effects studied under the second model specification.

Data analysis is implemented in Mplus (Muthén & Muthén, 1998–2011). In a SEM framework, since the underlying continuous variables are assumed to be jointly normally distributed, each pair of them (say y_i^* and y_j^*) is bivariate normal with correlation ρ_{ij}; these are known as the polychoric correlations (Jöreskog, 2005). Parameter estimation is done in three steps where thresholds are estimated in the first step from the univariate marginal distributions, and the polychoric correlations in the second from the bivariate distributions for given thresholds. In the third stage, the factor analysis model is fitted to the estimated polychoric correlation matrix using unweighted least squares (ULS), diagonally weighted least squares (DWLS), and weighted least squares (WLS). In WLS, the weight matrix is an estimate of the inverse of the asymptotic covariance matrix of polychoric correlations, while DWLS involves only the diagonal elements of that weight matrix. Recent studies confirm (Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009; Yang-Wallentin, Jöreskog, & Luo, 2010) that the WLS estimator converges very slowly to its asymptotic properties and therefore does not perform well in small sample sizes. DWLS and ULS are preferable to WLS and they seem to perform similarly well in finite samples. However, in order to compute correct standard errors and goodness-of-fit tests, the full weight matrix is needed. In our application, DWLS is used for estimation and WLS for obtaining the standard errors and test statistics.

The models being studied are the ones introduced in Section 2, with items y_{it}, $i = 1, 2, 3$, and the dropout indicators d_t used to give information on one attitudinal latent variable z_{a_t}

at waves $t = 1, \ldots, 5$ (with $d_1 = 0$ for all, as the observations are regarded as complete at the first wave). The latent variable captures attitudes towards women's work, with higher values of it indicating more conservative attitudes.

We first carried out two preliminary analyses, which allowed us to conclude that two assumptions introduced in Section 2 are satisfied in these data. First, we considered the assumption of measurement invariance, which states that each parameter of the measurement model (1)–(2) is the same at all time points t. A likelihood ratio test was carried out, and this constraint was not rejected against a model which allowed for non-invariance of measurement in the items. Next, for the second model specification we examined the assumption that the dropout latent variable z_d has its loadings λ_{d_t} set to 1 at all of $t = 2, \ldots, 5$, which for this model specication also implies that the attitudinal latent variable measured at the first wave (z_{a_1}) will have the same effect on dropout indicators at all time points. The model with this constraint was also not rejected against the unrestricted model where those loadings were allowed to vary freely across the time points.

Table 1 gives parameter estimates for the two model specifications along with their estimated standard errors (in brackets), when covariates are not yet considered. The attitude towards women's work loads very similarly on all three items, suggesting that the items contribute almost equally to measuring the attitude. The estimated thresholds for the dropout model are given in the second part of Table 1.

In the first model specification, the variance of the attitudinal latent variable at wave 1 is estimated as 0.32. The variance does not change much across waves, indicating that the variability of attitudes remains almost the same over time. The estimated covariance matrix of $\mathbf{z}_a$ for the first model specification is given by

$$\hat{\mathbf{\Gamma}} = \begin{bmatrix} 0.32 & 0.25 & 0.22 & 0.21 & 0.19 \\ & 0.34 & 0.26 & 0.24 & 0.22 \\ & & 0.34 & 0.26 & 0.24 \\ & & & 0.34 & 0.26 \\ & & & & 0.33 \end{bmatrix}.$$

The estimated covariances among the attitudinal latent variables are positive and significant, indicating a strong positive correlation of a person's attitude towards women's work across waves. As one would expect, the further apart the waves, the weaker is the covariance between the attitudes. Furthermore, a loading is estimated for each time-dependent attitudinal latent variable on the corresponding dropout indicator at the next wave. From Table 1, these loadings are negative and significant at 10% level of significance, indicating that the more conservative an individual's attitude is towards women's work, the less likely they are to drop out of the study at the next wave. The dropout is thus non-ignorable.

The last part of Table 1 gives results for the structural part of the second model specification. The estimated autoregressive parameter $\hat{\phi} = 0.874$, with estimated standard error of 0.007, again shows a significant and strong positive correlation of a person's attitude towards women's work over time. In other words, liberal/conservative views at a given wave are associated with liberal/conservative views at the preceding wave. The estimated dropout parameter $\hat{\beta} = -0.036$, with estimated standard error of 0.009, shows a significant dependence of dropout on attitude at the first wave, indicating non-ignorable dropout. The negative coefficient shows that the more conservative an individual's initial attitude is to-

Table 1: Parameter estimates for Models 1 and 2, for modeling attitudes towards women's work in the British Household Panel Survey.

		Model 1		Model 2	
		Measurement model			
		Est.	S.E.	Est.	S.E.
'Family'	λ_1	1		1	
'Contribution'	λ_2	1.115	(0.023)	1.115	(0.023)
'Independent'	λ_3	1.151	(0.025)	1.149	(0.025)
z_{a_1} on d_2^*	λ_{d2}	-0.014	(0.008)		
z_{a_2} on d_3^*	λ_{d3}	-0.019	(0.011)		
z_{a_3} on d_4^*	λ_{d4}	-0.044	(0.018)		
z_{a_4} on d_5^*	λ_{d5}	-0.056	(0.029)		
		Dropout model			
d_2^*	τ_{d2}	1.272	(0.022)	1.272	(0.022)
d_3^*	τ_{d3}	1.534	(0.027)	1.535	(0.027)
d_4^*	τ_{d4}	1.533	(0.028)	1.536	(0.028)
d_5^*	τ_{d5}	1.506	(0.029)	1.512	(0.029)
		Random effects			
Variances					
u_1	σ_{u1}^2	0.195	(0.007)	0.187	(0.007)
u_2	σ_{u2}^2	0.229	(0.008)	0.220	(0.008)
u_3	σ_{u3}^2	0.192	(0.008)	0.183	(0.008)
		Structural model			
Variance of z_{a_1}	σ_1^2	0.318	(0.011)	0.301	(0.010)
Autoregressive parameter	ϕ			0.874	(0.007)
Dropout parameter	β			-0.036	(0.009)

wards women's work, the less likely they are to drop out of the study. This conclusion too agrees with the one obtained from the first model specification. However, since the British Household Panel Survey is not a study of just women's work but also includes many other items (not analysed here), dropout is likely to be related to other factors as well.

The estimated means of the time-dependent attitudinal latent variable are, in order, 0.0, 0.057, 0.085, 0.101, and 0.103. This gradual increase in the mean indicates that as time goes by and people get older their views about women's work become more conservative. Another explanation is that since the more conservative people are less likely to drop out, the ones who remain in the study as time passes will tend to hold more conservative views.

The sample size considered here is large. In this situation, the X^2 goodness of fit statistic is not very helpful, as it will tend to suggest significant lack of fit even given very small discrepancies between the fitted and observed covariance matrices (Bijleveld et al., 1998). We therefore evaluate the two models by their Root Mean Square Error of Approximation (RMSEA) and Comparative Fit Index (CFI). The first model specification has an RMSEA of 0.017 and CFI 0.994, while the second has an RMSEA of 0.021 and CFI 0.991. The two model specifications seem to fit the data almost equally well, giving us the choice of which one to adopt. In this case, the second specification seems to be the more attractive option since it is more parsimonious and involves directed relationships rather than free correlations among the latent variables.

Next, three time-invariant covariates (sex as a dummy variable for women, age at first wave and initial educational attainment) and one time-varying covariate (occupational status) are introduced to the second model specification and allowed to affect both the attitude towards women's work at each wave and the dropout mechanism. Education is included as a binary variable that takes the value 1 if an individual has a medium or high academic qualification and 0 if no academic qualification is acquired. This is measured at the first wave and treated as time-invariant, as it tends to vary only slowly over time and is thus highly correlated across different waves. Occupational status is defined as a binary time-varying covariate which takes the value 1 if an individual is employed, retired or a student, and 0 if the individual is unemployed. The effect of covariates on the corresponding attitudes is constrained to be the same from wave 2 onwards. For the first wave, the effect of covariates on the attitude is allowed to be different, as this latent value is modeled solely as a function of covariates but not of previous attitudes.

Table 2 shows estimated regression coefficients of covariates on attitudes along with their estimated standard errors. Sex, initial age and education seem to have a significant effect on attitudes towards women's work at the first wave. The negative coefficient of sex indicates that, as expected, women seem to have more liberal attitudes towards women's work. Both age and education have significant positive coefficients on attitude at the first wave. This indicates that older people and people with at least medium or high education at the beginning of the study have more conservative views about women's work. This is in addition to the before-mentioned conclusion that as people get older (i.e. in the subsequent waves) their views tend to get still more conservative. Although occupational status does not seem to have a significant effect on attitude at the first wave, it does have a significant effect from wave 2 onwards, indicating that those who are employed, retired or students have more liberal attitudes towards women's work than the unemployed. Sex ceases to have a significant effect from wave 2 onwards. This is probably due to the fact that its effect is already carried through the attitude from previous waves.

Table 2: Parameter estimates for the regression of the attitudinal latent variables on covariates (sex, age, education and occupational status) for Model 2.

	Effect on z_{a_1}		Effect on $z_{a_2}, \ldots, z_{a_5}$	
	Est.	S.E.	Est.	S.E.
Sex (woman)	-0.049	(0.019)	-0.001	(0.006)
Age at first wave	0.001	(0.001)	-0.001	(0.000)
Education	0.197	(0.022)	0.026	(0.007)
Occupational status	-0.004	(0.032)	-0.104	(0.016)

Table 3 shows estimated regression coefficients of covariates measured at first wave on the dropout 'latent variable' z_d, along with their estimated standard errors. All the time-invariant covariates are significant. Sex has a negative coefficient, indicating that women are less likely to drop out. Age has a positive effect, meaning that older people are more likely to drop out, while the negative coefficient of education indicates that those with medium or high education are less likely to drop out of the study. In summary, older, less educated and male respondents have a higher propensity to drop out. It is worth mentioning that having accounted for those covariates, the dropout coefficient β of the attitude at the first wave is still significant, indicating nonrandom dropout. However, it is now positive (0.027), opposite to the coefficient in the model without covariates. Thus it now indicates that controlling for these covariates, the more conservative an individual is at the first wave, the more likely he or she is to drop out. The likeliest explanation of this reversal is controlling for education, for which higher education is associated with more conservative attitudes but also with lower probability of dropout.

Table 3: Parameter estimates for the regression of the dropout latent variable on covariates (sex, age, education and occupational status) for Model 2.

	Est.	S.E.
Sex (woman)	-0.110	(0.027)
Age at first wave	0.011	(0.001)
Education	-0.080	(0.032)
Occupational status	0.033	(0.050)

4. Conclusion

We have proposed two model specifications that incorporate dropout within the latent variable modeling framework to model multivariate longitudinal data. Both model specifications allow us to test whether the dropout depends on the variables of interest through the modeling of the probability of dropping out at a given wave as a function of the latent variables (in which case the dropout is nonrandom), observed covariates, or both latent variables and covariates. The models presented here are for ordinal observed variables and binary indicators for the dropout. Extensions to other types of observed variables are straightforward and do not require any further generalisations. The ordinal observed variables were modeled using underlying continuous variables and the classical factor analysis model, employing the three-step estimation procedure (thresholds, polychoric correlations, weighted least squares) as described in Jöreskog (1994, 2005). The dropout mechanism was modeled with a hazard function that may depend on the attitudinal latent variables and covariates. Different ways of modeling the relationships among the latent variables and the dropout mechanism were proposed and their advantages and disadvantages discussed. The proposed models remain within the standard framework of a general latent variable model for longitudinal data, and therefore estimation of model parameters and goodness-of-fit testing use conventional methods. Extensions of the proposed models to cope with intermittent missingness as well as item nonresponse are under development.

References

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: a unified approach.* London: Wiley.

Bijleveld, C. C. J. H., Mooijaart, A., van der Kamp, L. J. T., & van der Kloot, W. A. (1998). Longitudinal data analysis: Designs, models and methods. In L. J. T. van der Kamp & C. C. J. H. Bijleveld (Eds.), (p. 207-268). London: SAGE Publications Ltd.

Cagnone, S., Moustaki, I., & Vasdekis, V. (2009). Latent variable models for multivariate longitudinal ordinal responses. *British Journal of Mathematical and Statistical Psychology*, *62*, 401–415.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581–612.

Dantan, E., Proust-Lima, C., Letenneur, L., & Jacqmin-Gadda, H. (2008). Pattern mixture model and latent class model for the analysis of multivariate longitudinal data with informative dropouts. *The International Journal of Biostatistics*, *4*, 1-26.

Diggle, P. J., Heagerty, P. J., Liang, K., & Zeger, S. L. (2013). *Analysis of longitudinal data* (2nd ed.). Oxford: Oxford University Press.

Diggle, P. J., & Kenward, M. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics*, *43*, 49–93.

Dunson, D. (2003). Dynamic latent trait models for multidimensional longitudinal data. *Journal of the American Statistical Association*, *98*, 555–563.

Fieuws, S., & Verbeke, G. (2004). Joint modelling of multivariate longitudinal profiles: pitfalls of the random-effects approach. *Statistics in Medicine*, *23*, 3093–3104.

Fieuws, S., & Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, *62*, 424–431.

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 625-641.

Henderson, R., Diggle, P., & Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, *1*, 465–480.

Holman, R., & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*, 1–17.

Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G., & Mallinckrodt, C. (2006). Analyzing incomplete discrete longitudinal clinical trial data. *Statistical Science*, *21*, 52–69.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, *59*, 381–389.

Jöreskog, K. G. (2005). *Structural equation modelling with ordinal variables using LISREL* (Tech. Rep.). Scientific Software International.

Knott, M., Albanese, M. T., & Galbraith, J. (1990). Scoring attitudes to abortion. *The Statistician*, *40*, 217–223.

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

Molenberghs, G., Kenward, M. G., & Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika*, *84*, 33–44.

Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society, Series A*, *163*, 445–459.

Muthén, L. K., & Muthén, B. O. (1998–2011). *Mplus user's guide* (Sixth ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, B., Asparouhov, T., Hunter, A. M., & Leuchter, A. F. (2011). Growth modeling with nonignorable dropout: Alternative analyses of the STAR*D antidepressant trial. *Psychological Methods*, *16*, 17–33.

Muthén, B., & Masyn, K. (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral Statistics*, *30*, 27–58.

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: a latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society, Series A*, *162*, 177–194.

Roy, J. (2003). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics*, *59*, 829–836.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models.* Boca Raton, FL: Chapman & Hall/CRC.

Vasdekis, V., Cagnone, S., & Moustaki, I. (2012). A composite likelihood inference in latent variable models for ordinal longitudinal responses. *Psychometrika*, *77*, 425–441.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* New York: Springer-Verlag.

Wu, M. C., & Bailey, K. R. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics*, *45*, 939–955.

Wu, M. C., & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. *Biometrics*, *44*, 175–188.

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory Factor Analysis of ordinal variables with misspecified models. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*, 392-423.