Multivariate Longitudinal Data Subject to Dropout and Item Non-Response: A Latent Variable Approach

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Mai Hafez†

Department of Statistics, London School of Economics, UK

E-mail: m.m.hafez@lse.ac.uk†

Abstract

Longitudinal data are collected for studying changes across time. Studying many variables simultaneously across time (e.g. items from a questionnaire) is common when the interest is in measuring unobserved constructs such as democracy, happiness, fear of crime, social status, etc. The observed variables are used as indicators for the unobserved constructs "$latent\ variables$" of interest. Dropout is a common problem in longitudinal studies where subjects exit the study prematurely. Ignoring the dropout mechanism can lead to biased estimates, especially when the dropout is $non-ignorable$. Another possible type of missingness is item non-response where an individual chooses not to respond to a specific question. Our proposed approach uses latent variable models to capture the evolution of the latent phenomenon over time while accounting for dropout (possibly $non-random$), together with item non-response.

Introduction

- Longitudinal data are collected for studying changes across time.
- The problem of missing data is very common in longitudinal studies.
- One common type of missingness in longitudinal studies is $dropout$, where subjects often dropout of the study prematurely.
- Another type of missingness is item $non-response$ where an individual chooses not to respond to a specific question.

Why Latent Variable Models?

- In a longitudinal study, it is common that the phenomenon of interest is a latent "unobserved" construct or attitude.
- Latent variable models allow for the measurement of a latent construct or attitude using observed variables (items).
- One or few latent variables are assumed to explain correlations among the observed items, thus providing a means of dimension reduction as well as a summary of the unobserved attitude.
- When measuring changes in attitudes across time, several types of associations among the observed items need to be accounted for.
- Missingness can be incorporated within a latent variable model framework. Whether the missingness is random or not can also be tested within the model framework.

A Latent Variable Model for Multivariate Longitudinal Data

- A number of items is measured for each individual at each time point. An observed variable y_{it} denotes the i^{th} item at the t^{th} time point, where $i = 1, ..., p$ and $t = 1, ..., T$.
- At each time point, we assume there is a single continuous attitudinal latent variable z_{a_t} explaining correlations among the observed items.
- A random effect u_i is introduced for each observed item to account for the repetition of items over time.

Measurement Model

- Most variables in social surveys are measured on an ordinal scale.
- In a structural equation modelling (SEM) framework, ordinal items are treated using an underlying variable approach. It is assumed that each ordinal variable y_{it} is a manifestation of an underlying unobserved continuous variable y^*_{it}. For an ordinal variable y_{it} with c_{it} categories, the relationship between the ordinal and underlying variables is given in Jöreskog (2005) by

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

where

$$\tau^{(i)}_0 = -\infty, \quad \tau^{(i)}_1 < \tau^{(i)}_2 < \ldots < \tau^{(i)}_{c_{it}-1}, \quad \tau^{(i)}_{c_{it}} = \infty,$$

are known as thresholds.

- The underlying variable y^*_{it} is assumed to have a standard normal distribution. The measurement model for the underlying variables is now the classical factor analysis model

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

where λ_{it} is the loading of the latent variable z_{a_t} on the underlying variable y^*_{it}, u_i a random effect for the variable y^*_i, and ε_{it} an uncorrelated normally distributed random error.

Structural Model

- The time-dependent attitudinal latent variables z_{a_t} are assumed to be linked via a first-order autoregressive structure AR(1):

$$z_{a_t} = \alpha_t + \phi_a z_{a_{t-1}} + \boldsymbol{\theta'} \mathbf{x_t} + \delta_t, \quad t = 2, ..., T; \qquad (2)$$

where ϕ_a is a regression coefficient representing the dependence of the attitude at time t on that at the previous occasion $t-1$, δ_t is a random error assumed to have a normal distribution: $\delta_t \sim N(0, \upsilon_{\delta_t})$, and is uncorrelated with other errors or with $z_{a_{t-1}}$. $\mathbf{x}_t$ represents covariates and $\boldsymbol{\theta'}$ a vector of their corresponding coefficients.

- The joint distribution of the latent variables and random effects is given by

$$h(\mathbf{z_a}, \mathbf{u}) = h(\mathbf{z_a})\, g(\mathbf{u}) = h(\mathbf{z_a}) \prod_{i=1}^{p} g(u_i),$$

where $\mathbf{z_a} = (z_{a_1}, ..., z_{a_T})'$ is a $T \times 1$ vector of attitude latent variables, and $\mathbf{u} = (u_1, ..., u_p)$ is a $p \times 1$ vector of random effects.

- For identification purposes, the attitude latent variable at the first occasion z_{a_1} is normally distributed with mean 0 and variance σ^2_1.
- The latent variables will have a multivariate normal distribution $\mathbf{z_a} \sim MVN_{(T)}(\boldsymbol{\mu}, \Gamma)$ such that

$$\Gamma = \begin{bmatrix} \sigma^2_1 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1T} \\ & \sigma^2_2 & \sigma_{23} & & \sigma_{2T} \\ & & \sigma^2_3 & & \sigma_{3T} \\ & & & \ddots & \vdots \\ & & & & \sigma^2_T \end{bmatrix},$$

where Γ depends on the specification of the structural part of the model.

- The random effects are assumed to be normal $u_i \sim N(0, \sigma^2_{ui})$, $i = 1, ..., p$ and by orthogonality $g(\mathbf{u}) = \prod_{i=1}^{p} g(u_i)$.

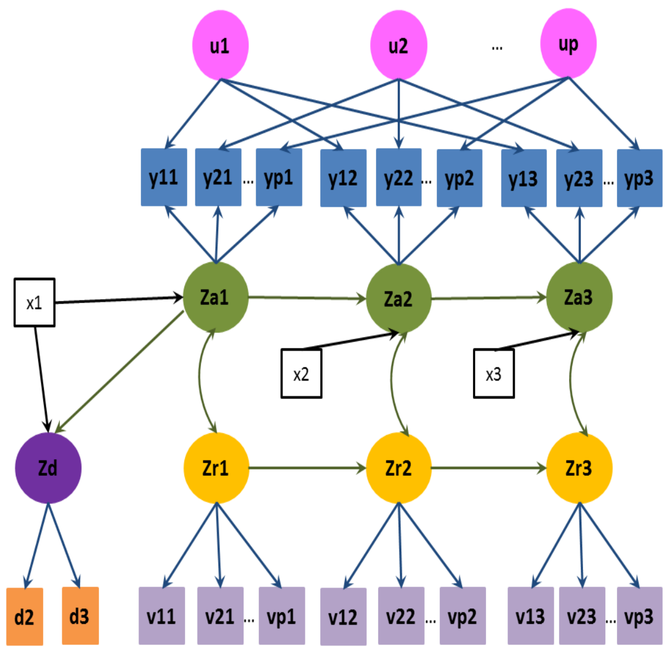

Figure 1: A path diagram for a model with three time points

Modelling The Missingness

Item Non-Response

- A pseudo-item v_{it} is created for each variable y_{it}, such that for a respondent who has not dropped out it takes the value 0 if a response is given, and 1 if the respondent does not respond to this item (O'Muircheartaigh and Moustaki (1999)):

$$v_{it} = \begin{cases} 0, & y_{it}\ is\ observed \\ 1, & y_{it}\ is\ not\ observed \end{cases}.$$

- If a respondent drops out, the value of pseudo-items for that respondent is recorded as "$missing$" .
- At each time point, a latent variable z_{r_t} representing response propensity is assumed to be underlying the pseudo-items.
- Each of the underlying continuous variables v^*_{it} is assumed to have a standard normal distribution and can be written as a function of the response propensity latent variable z_{r_t} as follows

$$v^*_{it} = \lambda_{v_{it}} z_{r_t} + \xi_{it}, \quad t = 2, ..., T, \qquad (3)$$

where $\lambda_{v_{it}}$ is the loading of the continuous latent variable z_{r_t} on the pseudo-items and ξ_{it} is an uncorrelated normal random error.

- Like the attitudes, the response propensity latent variables are assumed to be linked via a first-order autoregressive structure AR(1):

$$z_{r_t} = \alpha_{rt} + \phi_r z_{r_{t-1}} + \varrho_t, \quad t = 2, ..., T. \qquad (4)$$

- At any given time point, the attitudinal latent variable z_{a_t} is allowed to correlate freely with the corresponding response propensity latent variable z_{r_t}. If this correlation is significant, this can be taken as an indication of non-random missingness.

Dropout

- Dropout indicators $d_1, ..., d_T$ are defined as in Muthén & Masyn (2005) such that:

$$d_t = \begin{cases} 0, & \mathbf{y}_t\ is\ observed \\ 1, & \mathbf{y}_t\ is\ a\ dropout \end{cases},$$

where $\mathbf{y}_t = (y_{1t}, ..., y_{pt})'$ is a $p \times 1$ vector.

- Each of the underlying continuous variables d^*_t is assumed to have a standard normal distribution and can be written as a function of the "$dropout$" latent variable z_d as follows

$$d^*_t = \lambda_d z_d + e_t, \quad t = 2, ..., T, \qquad (5)$$

where λ_d is the loading of the continuous "$latent$" variable z_d on the dropout indicators and e_t is an uncorrelated normally distributed random error.

- It is assumed that the effect of the underlying latent variable z_d is the same on dropout indicators d_t at all time points.
- We assume that the dropout "$latent$" variable z_d has a normal distribution with both its mean and residual variance set to zero Muthén & Masyn (2005), which makes it a deterministic function of attitudes and covariates rather than a "$variable$". Defining z_d in such a way allows the dropout indicators to be affected by the attitudinal latent variables via z_d without them being directly considered as measures of the attitude;

$$z_d = \beta z_{a1} + \boldsymbol{\omega'_d} \mathbf{x_1} + \zeta. \qquad (6)$$

Data Analysis

- The data analysed here is from the British Household Panel Survey. A set of questions concerning women's role and views about women's work is addressed to participants on three waves (1993, 95, 97). The questions used as indicators of peoples' perceptions on women's work are:
  1. Woman and family happier if she works [Family]
  2. Husband and wife should both contribute [Contribution]
  3. Full-time job makes woman independent [Independent]
- Available responses are: strongly agree, agree, not agree/disagree, disagree and strongly disagree. The way the attitudinal latent variable is defined implies that the higher an individual scores on this latent variable, the more conservative their views are towards women's work.
- Data analysis is implemented in Mplus where estimation is done using weighted least squares using a diagonal weight matrix with standard errors that use a full weight matrix.
- The items analysed here have shown to be unidimensional. A chi-square difference test is performed, and measurement-invariance across time points is not rejected.
- The correlations between attitude towards women's work and response propensity are positive and significant at all time points indicating non-random missingness. More conservative attitudes are associated with lower propensity to respond.
- The dropout coefficient β is estimated by -0.08 (p-value = 0.134) indicating a random dropout at the 10 percent significance level.

References

[1] Little, R. and Rubin, D. (1987) *Statistical Analysis With Missing Data*. New York: John Wiley & Sons.

[2] Moustaki, I. , Jöreskog, K. G. and Mavridis, D. (2004) Factor models for ordinal variables with covariate effects on the manifest and latent variables: A comparison of LISREL and IRT approaches, *Structural Equation Modeling*, **11(4)**, 487-513.

[3] Muthén, B. and Masyn, K. (2005) Discrete-time survival mixture analysis.,*Journal of Educational and Behavioral Statistics*, **30(1)**,27-58.

[4] O'Muircheartaigh, C. and 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(2)**,177-194.