Suksma: Jurnal Psikologi Universitas Sanata Dharma
Vol. 1, No. 2, 36-45(ISSN: 1412-9426)

Multilevel Models with Binary/Ordinal Outcomes

Agung Santoso
Fakultas Psikologi, Universitas Sanata Dharma

Abstract. In the current article, three multilevel models involving binary or ordinary outcomes were reviewed. The three models proposed to solve problems related to using dichotomous or ordinal variables in multilevel models. Reviews on and comparison between the three models were presented in the current article. The article was then concluded by discussion on the limitation of the three models and future direction for research.

Keywords: Multilevel model, longitudinal analysis, binary outcome, graded response model, ordinal outcome

Introduction

Multilevel model has become increasingly popular in psychological research. Its popularity is due to the model capability to estimate not only fixed effects but also random effects. Multilevel model also takes into account the dependency of observations collected in each cluster at each level. The model can be employed when the data are hierarchical in nature, including data from longitudinal studies. In such studies, observations across time, as the first level, are clustered in each person, as the second level.

Many researchers have developed multilevel models to address some of its limitations. Two of such limitations are assumptions that outcome variables are continuous and measured without error, while psychological data are often binary / ordinal in nature and not error-free. Application of regular multilevel model that ignores the ordinal nature of the data is not appropriate because it may result in out of bound predicted values and inflated estimates of random slope variances and cross-level interaction (Bauer & Sterba, 2011). Measurement error that is not accounted for in the model may make the standard errors of parameters underestimated (Fox, 2007).

Korespondensi Penulis
Agung Santoso, Fakultas Psikologi Universitas Sanata Dharma, Yogyakarta.
Email: *agungsan@usd.ac.id*

$$\begin{aligned}
\text{Level 1: } & Y_{ij} = \beta_{0j} + \beta_{1j}X_{1j} + r_{ij} \\
\text{Level 2: } & \beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \\
& \beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j} \\
& \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{10} \\ \tau_{01} & \tau_{11} \end{bmatrix}, \right)
\end{aligned}$$

The needs to estimate growth trajectory of psychological constructs, such as cognitive abilities, across time using different measures may also demand an enhanced multilevel modeling. The use of different measures is often based on justified rationale, such as age appropriateness and improved test batteries over time. However, such condition may complicate the efforts to separate differences in changing scales over time from changes in the constructs over time. Therefore, a model is needed to relax the requirements of using the same measures over time while the model still enables researchers to establish a common scale so that changes in the constructs over time can be estimated.

Current paper is focused on three models that may address concerns mentioned above. The models are presented in a sequential manner; begin with multilevel cumulative probit / logit model, longitudinal invariant Rasch test, and longitudinal item response theory-distribution parameter estimation. In the discussion section, comparison of models, including their advantages and limitations were discussed. Current paper is concluded by proposing several directions for future studies.

Multilevel Cumulative Probit / Logit Model

Bauer and Sterba (2011) proposed a model called multilevel cumulative probit/logit model to solve modeling with ordinal nature of the outcome variables. The author use two ways of approaching the model by: (a) conceiving the ordinal outcomes as a coarse categorization of an underlying continuous variable, and (b) using a multinomial distribution to specify the conditional distribution of the ordinal outcome variable.

Let Y_{ij} be the underlying continuous outcome variable for person i in cluster j; β_{0j} and β_{1j} be, respectively, the intercept and slope parameters, while X_{ij} be the explanatory variable at level 1; $\gamma_{00}, \gamma_{01}, W_j$ be the intercept and slope parameters, and explanatory variable of β_{0j}, respectively; $\gamma_{10}, \gamma_{11}, W_j$ be the intercept and slope parameters, and explanatory variable of β_{1j}; and r_{ij}, u_{0j}, and u_{1j} be the error terms. The first approach to the model can be expressed as the following:

In the model, r_{ij} can be assumed to follow either the standard normal distribution, $N(0,1)$, that lead to multilevel cumulative probit model, or a logistic distribution, $logistic\left(0,\frac{\pi^2}{3}\right)$, that lead to multilevel logit model. Note that the variance of the model is fixed.

A threshold model was also posited by the authors to link Y_{ij} to the ordinal outcome variable:

$Y_{ij} = 1 if Y_{ij} < \nu^{(1)}$

$Y_{ij} = 2 if \nu^{(1)} \leq Y_{ij} < \nu^{(2)}$

...

$Y_{ij} = C if \nu^{(C-1)} \leq Y_{ij}$

where Y_{ij} is the observed ordinal outcome variable and $\nu^{(c)}$ is a strictly increasing threshold parameter (i.e. $\nu^{(1)} < \nu^{(2)} < \cdots < \nu^{(C-1)}$). This threshold model shows that if a person's observed score on the outcome variable are at *c* category, his score on Y_{ij} must have passed the threshold for that category.

A different approach can attain equivalent models by using the generalized linear model framework. In this framework, Y_{ij} is assumed to follow a multinomial distribution, with parameters describing the probabilities of the categorical responses. Cumulative coding variables, $Y_{ij}^{(c)}$, are defined to capture the ordered nature of the categories of the observed outcome. An amount of $C-1$ coding variables are defined such that $Y_{ij}^{(c)} = 1$ if $Y_{ij} \leq c$. The coding variables for the last category, $Y_{ij}^{(C)}$, is omitted since it is always scored 1 for all Y_{ij}.

The second approach of the model can be expressed as the following

$$\begin{aligned} \text{Level 1: } & \eta_{ij} = \beta_{0j} + \beta_{1j}X_{1j} \\ \text{Level 2: } & \beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \\ & \beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j} \end{aligned} \quad (1)$$

η_{ij} is treated as person's score on latent-outcome-continuous variable, like ability level in item response model. The model for observe response is given as the following:

$$Y_{ij}^{(c)} = g^{-1}\left[\nu^{(c)} - \eta_{ij}\right] + r_{ij} \quad (2)$$

where $\nu^{(c)}$ is the threshold parameter for category *c* that allows for increasing probabilities that is accumulated across categories, and $g^{-1}[.]$ is the inverse link function, a function that maps a continuous range of $\left[\nu^{(c)} - \eta_{ij}\right]$ to the bounded zero to one range. We can choose any function that has zero and one asymptotes to be $g^{-1}[.]$. Two functions commonly chosen are the standardized normal CDF and inverse logistic function.

Combining (1) and (2) provides

$$\text{Level 1: } Y_{ij}^{(c)} = g^{-1}\left[v^{(c)} - \beta_{0j} - \beta_{1j}X_{ij}\right] + r_{ij} \quad (3)$$

If $g^{-1}[.]$ is an inverse logistic function, then

$$\text{Level 1: } Y_{ij}^{(c)} = \frac{\exp(v^{(c)} - \beta_{0j} - \beta_{1j}X_{ij})}{1 + \exp(v^{(c)} - \beta_{0j} - \beta_{1j}X_{ij})} + r_{ij} \quad (4)$$

The authors posited two other constraints to the model to make it identified. The first constraints were related to thresholds and overall model intercepts. Because the thresholds and intercepts cannot be estimated jointly, one should either set the intercept to be zero to estimate all thresholds or set the first threshold to be zero and estimate the intercept. The authors chose the latter constraint because it is the most common practice. The second constraint relates to the coefficients in level 1, which are set to be the same across all categories.

Two procedures were chosen to estimate the parameters: the penalized quasi-likelihood (PQL) and maximum likelihood with adaptive quadrature approach (ML). The authors compared the performance of the two procedures in estimating the parameters. The results showed that although in many conditions PQL provided negatively biased estimates, the estimates had smaller MSE compared to the estimates from using ML. The small MSE means that PQL estimates had biases that are compensated by small variability across samples.

Their study also showed that using the regular multilevel model by assuming continuous outcome variables provided negatively biased estimates for random effects and their dispersion estimates. Only when the distribution of the response categories were roughly normal and the number of categories was seven, the regular model provided a tolerable size of bias.

Longitudinal Invariant Rasch Test

This model is proposed by McArdle, Grimm, Hamagami & Bowles (2009) to address the need to estimate growth-decline trajectory of intellectual abilities from datasets from three studies. The datasets were taken from 419 participants measured repeatedly (T=16) starting when they were 2 years old and ended when they were 72 years old. Several tests were employed to match the age of the participants with the age required by the test. Two objectives were to be attained in this article: testing the longitudinal measurement invariance / equivalence and estimating growth-decline trajectory.

The longitudinal invariant Rasch test (LIRT) model can be seen as an extension of growth of factor scores model, in which the measurement level is posited as the first level, while the first and second level of the longitudinal analysis are posited as the second and third level, respectively. The model for each level can expressed as the following:

$$\begin{aligned}
\text{Measurement Level: } & \ln\left(\frac{Pr[t]_{i,n}}{1+Pr[t]_{i,n}}\right) = g[t]_n - \beta_i \\
\text{Level 1: } & g[t]_n = g_{0n} + A[t]g_{1n} + u[t]_n \\
\text{Level 2: } & g_{0n} = \nu_{00} + \nu_{01}X_n + d_{0n} \\
& g_{1n} = \nu_{10} + \nu_{11}X_n + d1n \\
& \begin{bmatrix} g_{0n} \\ g_{1n} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \begin{bmatrix} \phi_0^2 & \phi_{01} \\ \phi_{10} & \phi \end{bmatrix}\right)
\end{aligned}$$

where $Pr[t]_{i,n}$ is the probability of answering item *i* correctly by *n* at time *t*; $g[t]_n$ is the ability level of person *n* at time *t*; β_i is the difficulty parameter of item *i*; g_{0n} is individual's initial level; g_{1n} is the individual's slope of change over time; $A[t]$ is time or shape of change over time; $u[t]_n$ is time specific unique score. At level two, ν_{00} and ν_{10} are the intercepts aggregating individuals' intercepts and slopes, respectively; ν_{01} and ν_{11} are the slopes representing relationships between observed predictor X_n and the individuals' intercepts and slopes, respectively; while d_{0n} and d_{1n} are the error terms.

In this model, $A[t]$ can be defined in various ways to answer a specific questions posed by researchers. For example, in this study, because McArdle, et.al (2009) wanted to estimate growth-decline trajectory, they defined $A[t] = exp(-\pi_g Age[t]) - exp(-\pi_d Age[t])$, where $\pi_g$is the rate of growth and π_d is the rate of decline.

The authors conducted two ways of estimating LIRT parameters: (a) two-stages approach, and (b) simultaneous approach. In the two-stages approach, the ability and difficulty parameters were estimated for each occasion and then used the ability estimates as the outcome variable in the longitudinal analysis part. Such approach ignores dependencies of within person ability estimate across time that may result in biased and inefficient estimates. In the simultaneous approach, the ability and difficulty parameters and parameters of the longitudinal model are estimated simultaneously that may provide more efficient estimates and more precise hypothesis tests.

MLE-MAR estimation procedure, with integration method using Gauss-Hermitte quadrature, were employed for two-stages and simultaneous approach due to incomplete data. The authors also employed MCMC to estimate simultaneous approach for its computing efficiency. The MLE-MAR estimation was conducted using NLME package in SAS while MCMC estimation was conducted using WinBUGS. Several prior distributions used to conduct MCMC were as the following:

$$\begin{aligned}
\beta_i &\sim \mathcal{N}(0, 10^{-6}) \\
\sigma_{u[t]_n} &\sim \mathcal{IG}(0.001, 0.001) \\
\mu_0 &\sim 0 \\
\mu_1 &\sim \mathcal{N}(0, 10^{-6}) \\
\Phi &\sim \mathcal{IW} \\
\pi_d &\sim \mathcal{N}(0, 10^{-6}) \\
\pi_g &\sim \mathcal{N}(0, 10^{-6})
\end{aligned}$$

Estimation using MLE-MAR, particularly for simultaneous approach, took a far longer time than expected in conducting the analysis, while MCMC provided estimation in a more reasonable time required for conducting the analysis.

Based on the results, the authors compared the two approaches of estimating LIRT parameters. The comparisons can be seen in Table 1.

Table 1. A Comparison of Two Stage Approach and Simultaneous Estimation

Two Stage Approach	Simultaneous Estimation
Advantages	
- More practical and intuitive - Immediately provide additional information such as ability plot across time - Evaluation of Item and Person fit-ness can be conducted - Outliers detection of ability estimates - Computationally efficient	- Taking into account dependency within person across time - Unbiased and more efficient estimates
Disadvantages	
- Not accounted for dependency within person across time - May result in biased and / or less efficient estimates	- Computationally cumbersome

Longitudinal Item Response Theory-Distribution Parameter Estimation

Andrade & Tavares (2005) proposed another approach to estimate parameters in longitudinal data acquired using different tests without estimating ability parameters of each person. The approach requires that researchers have known the item parameters beforehand. The authors claimed that their approach works even without overlapped items across tests across time.

The authors proposed the model as the following:

$$\text{Item Level: } P_{jit} = P(U_{jit} = 1|\theta_{jt}, \zeta_i)$$

$$\text{Test / Time Level: } P(U_{jit}|\theta_{jt}, \zeta_i) = \prod_{i \in I_t} P(U_{jit}|\theta_t, \zeta_i)$$

$$\text{Person Level across Test / Time: } P(U_{j.}|\theta_{jt}, \zeta) = \prod_{t=1}^{T} \prod_{i \in I_t} P(U_{jit}|\theta_t, \zeta_i)$$

where U_{jit} is binary response by person j on item i in test t, where $j = 1,2,\dots,N$, $i = 1,2,\dots,n$ and $t = 1,2,\dots,T$; θ_{jt} is the ability level of person j measured by test t; ζ_i is the matrix of known item parameters either from one, two or three parameter logistic model in test t; $U_{j.t} = \left(U_{j1t}, U_{j2t}, \dots, U_{jn_t t}\right)'$;$U_{j..} = \left(U'_{j.1}, U'_{j.2}, \dots, U'_{j.T}\right)$; $\zeta = (\zeta'_1, \zeta'_2, \dots, \zeta'_n)$;$I_t$ is a set of items numbers that is used in test t so that $n \leq n_c = \sum_{t=1}^{T} n_t$; and $\theta = (\theta_1, \theta_2, \dots, \theta_T)$.

By assuming that θ follows a multivariate normal distribution with parameters η, the unconditional probability of pattern $U_{j..}$ can be expressed as the following:

$$P(U_{j.}|\theta_{jt}, \zeta) = \int_{R^T} P(U_{j..}|\theta, \zeta) g(\theta|\eta) d\theta \tag{5}$$

where $g(\theta|\eta)$ is the density function of the multivariate normal distribution. The above probability depends on the known item parameters and η, the parameters of θ distribution.

The likelihood equation is derived from the probability of observing a certain response pattern $U_{j..}$ given item parameters ζ and distribution parameters η, which follows multinomial distribution given by

$$P(R|\zeta, \eta) = \frac{N!}{\prod_{j=1}^{s} r_j!} \prod_{j=1}^{s} [P(U_{j..}|\zeta, \eta)]^r \tag{6}$$

where j is now represents one of the s different response pattern where $s \leq min(N, 2^{n_c})$, so that $j = 1,2,\dots,s$; $R = (r_1, r_2, \dots, r_s)'$ is the $(s \times 1)$ vector of frequencies of observing response pattern $U_{j..}$. Taking the first derivative of the log of (6) provides:

$$\sum_{j=1}^{s} r_j \int_{R^T} \left(\frac{\partial}{\partial \eta} \log g(\theta|\eta) \right) g_j^*(\theta) d\theta \tag{7}$$

where

$$g_j^*(\theta) = \frac{P(\theta_{j.}|\theta, \zeta) g(\theta|\eta)}{P(U_{j.}|\zeta, \eta)} \tag{8}$$

The authors implementing the model by using multivariate normal distribution as $g(\theta \vee \eta)$ with different type of covariance structures like diagonal, uniform, banded, heterogenous, and unstructured covariance matrix. The selection of the types of covariance structure is based on theoretical consideration as well as information from pilot study. Estimation was then conducted using MLE with Newton-Rhapson method for numerical analysis.

Discussion and Future Direction

The three models have previously been reviewed independently to show how the model were defined and estimated. Here, comparison between models is given in Table 2 (see *Attachment 1*), to

show their similarities, relative advantages, and limitations. The differences between the three models are actually not substantial. One can apply some features of one model to another model. For example, one can extend LIRT model to incorporate not only binary responses but also ordinal responses or graded responses. One can also apply different types of covariance structure from LIRT-DPE model in LIRT or MCPL model. This means that, in the future, one may construct a more general model that may cover all three models that potentially may offer more useful-features.

For example, the model may also address one limitation that applies to all three models, that is exclusion of measurement model in explanatory variable. Therefore, in the model, one can extend the idea of applying IRT not only on outcome variable but also on explanatory variables. The complexity of the model may also cause problems in terms of computational efficiency. Therefore, in the future, one can search for either estimation (e.g. GEE) or computational methods (i.e. integration or numerical analysis) that are more efficient and less burdensome.

References

Andrade, D. F., & Tavares, H. R. (2005). Item response theory for longitudinal data: population parameter estimation. *Journal of Multivariate Analysis*, *95*(1), 1–22. https://doi.org/10.1016/j.jmva.2004.07.005

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods*, *16*(4), 373–390. https://doi.org/10.1037/a0025813

Fox, J.-P. (2007). Multilevel IRT modeling in practice with the package mlirt. *Journal of Statistical Software*, *20*(5), 1 – 16. https://doi.org/10.18637/jss.v020.i05

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods*, *14*(2), 126–149. https://doi.org/10.1037/a0015857

Attachment 1

Table 2. Comparison between model on several issues

	Model		
	MCPL	**LIRT**	**LIRT-DPE**
Estimation	- Penalized quasi-likelihood - Maximum likelihood using quadrature integration	- MLE using Gauss-Hermitte quadrature integration, computationally burdensome - MCMC	- MLE using Newton Rhapson for numerical analysis - Item parameters need to be known beforehand
Ordinal / Binary outcome	- Ordinal and Binary outcome can be addressed - One Parameter Logistic / Probit model (1PL)	- Binary outcome, extendable to ordinal outcome - 1PL, extendable to graded response and 2 or 3PL	- Binary outcome, extendable to ordinal outcome - 1PL, extendable to 2 or 3 PL
Measurement error of outcome	- Assume no measurement error	- Measurement error is modeled	- Measurement error is modeled
Ability estimate	- Based on observed outcome	- Ability estimate can be obtained	- Ability estimate cannot be obtained

Growth trajectory	- Growth trajectory can be obtained	- Growth and decline parameters can be obtained - More complex definition of 'time'	- Only mean and covariance estimate, can be extended to estimate growth trajectory
Variance-Covariance structure	- 'regular' covariance structure, fixed variance for first level residuals	- 'regular' covariance structure	- Several types of covariance structure
Measurement error of explanatory	- Assume no measurement error	- Assume no measurement error	- No explanatory variable on second level
Longitudinal measurement invariance	- NA (multilevel model)	- Longitudinal measurement invariance can be tested with overlapping items	- Assume longitudinal measurement invariance, but cannot be tested using the model

Note: MCPL = Multilevel Cumulative Probit/Logit, LIRT = Longitudinal Invariant Rasch Test, LIRT-DPE = Longitudinal Item Reponse Theory-Distribution Parameter Estimation.