Modeling Interactions Between Latent and Observed Continuous Variables Using Maximum-Likelihood Estimation In Mplus

Bengt Muthén
University of California, Los Angeles

Tihomir Asparouhov
Muthén & Muthén *

Mplus Web Notes: No. 6
Version 1, March 5, 2003

*The research of the first author was supported under grant K02 AA 00230 from NIAAA. The email address of the first author is bmuthen@ucla.edu.

Abstract

Modeling with random slopes is used in random coefficient regression, multilevel regression, and growth modeling. Random slopes can be seen as continuous latent variables. Recently, a flexible modeling framework has been implemented in the Mplus program to do modeling with such latent variables combined with modeling of psychometric constructs, typically referred to as factors, measured by multiple indicators. This note shows how such a framework can handle interactions between latent continuous and observed continuous indicators. Three examples are given: a Monte Carlo simulation to estimate power to detect the interaction; a psychological example; and a growth modeling example. Mplus input, output, and data are available at the Mplus web site, www.statmodel.com/mplus/examples/webnote.html.

1 Introduction

Modeling with random slopes is used in random coefficient regression, multilevel regression, and growth modeling. Random slopes can be seen as continuous latent variables. Recently, a flexible modeling framework has been implemented in the Mplus program to do modeling with such latent variables combined with modeling of psychometric constructs, typically referred to as factors, measured by multiple indicators. This note shows how such a framework can handle interactions between latent continuous and observed continuous indicators.

It is useful to first review the types of interactions that are of interest. The focus is on the standard situation with interactions between variables that are exogenous as opposed to endogenous. The standard case concerns two continuous observed variables interacting in their influence on a dependent variable (see Aiken & West, 1991), where the interaction is handled by including a product of the two variables in a regular linear regression. The following additional interaction cases may be considered.

Case A: Interaction between a latent continuous variable and an observed categorical variable

This type of interaction is handled by conventional structural equation modeling (SEM) using multiple-group analysis, where the observed (unordered) categorical variable represents the groups. The regression parameter for the latent continuous variable predicting a dependent variable can vary across the groups. For example, Muthén and Curran (1997) considered 2-group growth modeling to assess treatment effects in a randomized preventive intervention, where the model included a regression of the growth rate on the initial status and where this regression was different for treatment and control group individuals.

Case B: Interaction between a latent categorical variable and an observed variable

This type of interaction can be handled by mixture modeling as implemented in Mplus (see Muthén, 2002). The latent categorical variable represents the latent classes of the mixture modeling. The observed variable, be it continuous or categorical, can have different influence on a dependent variable for different latent classes. For example, in the growth mixture analyses of Muthén (2001) variation across trajectory classes was allowed for the influence of time-invariant covariates on the intercept and slope in a growth model.

Case C: Interaction between a latent continuous variable and an observed continuous variable

This type of interaction cannot be handled by conventional SEM. Special interaction modeling involving latent variables is needed, e.g. using the Joreskog-Yang approach (Joreskog & Yang, 1996), 2SLS (Bollen, 1996), or the full-information maximum-likelihood approach of Klein and Moosbrugger (2000).

Case D: Interaction between latent continuous variables

Same comment as for Case C.

This note considers case C. It will be shown that this fits into the Mplus latent variable framework so that full-information maximum-likelihood estimation is possible (Asparouhov & Muthén, 2002). Klein and Moosbrugger (2000) pointed out the important efficiency and power advantages for interaction modeling by use of full-information maximum-likelihood estimation as compared to limited-information estimators such as the Joreskog-Yang and 2SLS approaches.

The Muthén Curran (1997) and Muthén (2001) growth examples mentioned above indicate that interaction modeling involving latent variables is of interest not only in cross-sectional data with psychometric factors, but also in growth modeling where there are not necessarily any psychometric factors. In growth modeling, the random effects of initial status and growth rate are latent continuous variables that may interact with observed variables. Seltzer, Choi, and Thum (2002a, b) and Thum (2002) extended the Muthén-Curran regression of the growth rate on the initial status to allow for interaction between the initial status and time-invariant covariates. The approach to be discussed here is applicable to this type of modeling as well.

Section 2 describes the random slope approach to interaction modeling. Section 3 considers a Monte Carlo simulation study to show how power to detect interaction effects can be studied. Section 4 considers a cross-sectional data example. Section 5 considers a growth modeling example. Section 6 concludes. Mplus input, data, and output for the analyses in this note are available at www.statmodel.com/mplus/examples/webnote.html For explanation of Mplus commands, see Muthén and Muthén (1998-2001) and also the new random slope features in Muthén and Muthén (2003).

2 Random Slopes Modeling of Interactions With A Latent Continuous Variable

Consider for simplicity a single continuous observed dependent variable y observed for individual i,

$$y_i = \beta_0 + \beta_1\, x_i + \beta_2\, \eta_i + \beta_3\, \eta_i\, x_i + \epsilon_i, \tag{1}$$

where x_i is an observed covariate, η_i is a latent continuous variable (factor) measured with multiple indicators $y_1, \ldots, y_p$,

$$y_{ij} = \nu_j + \lambda_j\, \eta_i + \epsilon_{ij}, \tag{2}$$

using a regular factor-analytic measurement model.

When an asymmetric interpretation of the η, x interaction in (1) is of substantive interest, it is useful to rearrange (1) as the equivalent expression

$$y_i = \beta_0 + \beta_2\, \eta_i + (\beta_1 + \beta_3\, \eta_i)\, x_i + \epsilon_i. \tag{3}$$

This uses the "moderator function" $(\beta_1 + \beta_3\, \eta_i)$[1], such that η_i moderates the influence of x on y. The interpretation of the moderator function is aided by considering β coefficients corresponding to standardized η and x variables. This can be achieved by centering of x and parameterizing η to have zero mean, followed by a multiplication of β coefficient estimates by the estimated standard deviations for x and η. Note that unlike conventional SEM, the variance of y_i conditional on x_i changes as a function of x_i.

The Mplus approach to handling the interaction in (1) is as follows. The equation in (1) can be rewritten using two equations involving a random slope variable s_i,

$$y_i = \beta_0 + \beta_1\, x_i + \beta_2\, \eta_i + s_i\, x_i + \epsilon_i, \tag{4}$$

$$s_i = 0 + \beta_3\, \eta_i + 0. \tag{5}$$

In (4), a random slope is defined for the observed covariate x_i. In (5), the random slope is taken to be the same as the factor η_i, except for a scaling factor β_3, the interaction coefficient that we want to estimate. In (5), s_i is a latent variable that only contributes a single additional parameter, β_3. The two equations of (4), (5) indicate that the model is specified as a general SEM extended to random slopes. Such random slopes models can be handled in Mplus and three examples are discussed in this note.

The model in (4), (5) can be generalized in line with SEM, for example replacing y_i in (4) with a latent continuous dependent variable having multiple indicators. Several interaction terms may be considered. Muthén and Asparouhov (2003) discuss multilevel modeling with a cross-level interaction between a level 1 covariate and a level 2 factor.

Muthén and Asparouhov (2002) give an example of using random slopes to model heteroscedastic residual variances. Such heteroscedasticity modeling can be combined with interaction modeling.

[1]We thank Andreas Klein for suggesting this representation.

3 Example 1: A Monte Carlo Simulation

Monte Carlo simulation can be used to study the behavior of estimators in terms of parameter estimate bias, standard error performance, coverage, and power (for an introductory description of Mplus Monte Carlo simulations, see, e.g. Muthén & Muthén, 2002). Of particular interest with interaction modeling is the power to reject that the interaction is zero.

Consider the following model with an interaction between a factor η_2 and an observed continuous covariate x,

$$y_{1i} = \nu_1 + \eta_{1i} + \epsilon_{1i}, \tag{6}$$

$$y_{2i} = \nu_2 + \lambda_{21}\, \eta_{1i} + \epsilon_{2i}, \tag{7}$$

$$y_{3i} = \nu_3 + \eta_{2i} + \epsilon_{3i}, \tag{8}$$

$$y_{4i} = \nu_4 + \lambda_{42}\, \eta_{2i} + \epsilon_{4i}, \tag{9}$$

$$\eta_{1i} = \beta_1\, x_i + \beta_2\, \eta_{2i} + \beta_3\, \eta_{2i}\, x_i + \zeta_i, \tag{10}$$

with the usual assumptions on the residuals and with zero mean for η_2. Data were generated from this model using a normally distributed x and normally distributed residuals. Parameter values were chosen to give indicator reliabilities of 0.80. 500 Monte Carlo replications were used. Different sample sizes were tried. The Mplus output for Example 1 in the Mplus Web Note section uses $n = 400$. The Mplus output shows the average parameter estimates, the standard deviations of the estimates, the average s.e.'s, the 95% coverage, and the proportion of replications for which the hypothesis of a parameter being different from zero was rejected (estimated power).

It is seen from the output that the behavior of the maximum-likelihood estimator is very good, with excellent coverage close to 95%. At this sample size the power to detect the interaction effect is estimated as 0.89. With $n = 300$ (not shown) the power goes down to 0.78. For $n = 300$, however, 7 of the 500 replications resulted in a non-admissible solution with zero variance estimates for the indicators of η_2. For $n = 200$ there were 25 non-admissible solutions with an estimated power of 0.61. The non-admissible solutions may be due to the relatively small sample size, relatively high reliability, and having only two indicators per factor.

Further Monte Carlo investigations in Mplus might include the impact of model misspecification. For example, data generated with an interaction can be analyzed without it to see how much misestimation is produced.

4 Example 2: A Psychological Study Of Behavior

The model given in the previous section was also used for real-data analysis.[2] The analysis considers a sample of $n = 156$ mothers, where η_1 is mother's harsh discipline, η_2 is boy disruptive behavior, and x is early difficult temperament of boy.[3]

The Mplus output for Example 2 in the Mplus Web Note section shows that the interaction just reaches significance and is positive. This implies that higher than average early difficult temperament of boy and higher than average boy disruptive behavior interact to produce higher mother's harsh discipline.

In this analysis, two residual variances approach zero and are fixed at zero. The zero residual variances are not unexpected in light of the Monte Carlo simulations in the previous section, given the relatively small sample size. The maximum-likelihood estimates produced by Mplus are close to those of the LMS method of Klein and Moosbrugger (2000).

5 Example 3: Growth Modeling With Intercept-Covariate Interaction

Consider a conventional linear growth model with random intercepts η_0 and random slopes η_1,

$$y_{it} = \eta_{0i} + \eta_{1i}\, a_{it} + \epsilon_{it}, \tag{11}$$

where a_{it} is a time-related variable, e.g. scored as $a_{i1} = 0$ to define the intercept as the initial status of the growth model. Consider the extension of this model to regression of the slope on the intercept as well as on the interaction between the intercept and a time-invariant covariate x,

$$\eta_{1i} = \beta_0 + \beta_1\, \eta_{0i} + \beta_2\, x_i + \beta_3\, \eta_{0i}\, x_i + \zeta_i, \tag{12}$$

where η_0 and x are allowed to be correlated (alternatively η_0 can be regressed on x).

This model is applied to data from the Longitudinal Study of Youth (LSAY), which is a national sample of mathematics and science achievement of students in US public schools. Data are considered on mathematics achievement for grades 7-10 and is available at the Mplus Web Note section. A covariate representing math courses taken in grade 7 is considered.[4] The grade 7 course taking covariate is used as a predictor of the slope,

[2] A minor modification is estimating the loading for y_3 on η_2 while fixing the η_2 variance at 1 in order to have standardization already accomplished.

[3] We thank Mike Stoolmiller and Deborah Capaldi for making this example available.

[4] This covariate concerns the highest math course taken during grade 7, using the scoring 0 = no course, 1 = low,basic, 2 = average, 3 = high, 4 = pre-algebra, 5 = algebra I, 6 = geometry, 7 = algebra II, 8 = pre-calc, 9 = calculus.

both as a main effect and as an interaction with the initial status in grade 7. A key idea is that the initial status (η_0) of a student influences the ability to benefit from instruction and therefore influences the growth rate (η_1) in developing mathematics skills. In addition, the grade 7 course taking and initial status may interact in their influence on the achievement growth rate.

The Mplus output for Example 3 in the Mplus Web Note section shows a significant negative interaction between the initial status factor η_0 and the course taking covariate x.[5] Here, the moderator function interpretation discussed in Section 2 may be applied given that initial status may be seen as moderating the influence of the course taking in grade 7. The estimated right-hand-side of (12), standardized to unit variances for the η_0, x predictors, is then rearranged as

$$\hat{\beta}_0 + \hat{\beta}_1\,\eta_0 + (\hat{\beta}_2 + \hat{\beta}_3\,\eta_{0i})\,x = 0.417 + 0.080\,\eta_0 + (0.041 - 0.039\,\eta_{0i})\,x. \qquad (13)$$

The interpretation is as follows. At one standard deviation above the initial status mean ($\eta_0 = 1$), the slope increases only 0.002 ($0.041 - 0.039$) for a standard deviation increase in course taking. At the mean of initial status ($\eta_0 = 0$), the slope increases 0.041 for a standard deviation increase in course taking. At one standard deviation below the initial status mean, the slope increases 0.080 ($0.041 + 0.039$) for a standard deviation increase in course taking. Simply put, seventh grade math course taking matters for low starters, but not for high starters.

6 Conclusion

This note has shown that the Mplus capability of latent variable modeling with a combination of random slopes and factors makes it possible to handle interactions between latent and observed continuous variables. This Mplus feature allows for full-information maximum-likelihood estimation. As pointed out in Klein and Moosbrugger (2000), the ability to use full information maximum-likelihood estimation as opposed to the limited-information approaches typically used for interaction modeling in SEM gives an important efficiency and power advantage.

[5] To aid the interpretation, the course taking covariate is centered at its mean and the initial status factor η_0 is parameterized to have zero mean by adding time-invariant intercepts ν to (11).

References

Aiken, L.S. & West, S.G. (1991). *Multiple Regression: Testing and Interpreting Interactions.* Newbury Park: Sage Publications.

Asparouhov, T. & Muthén, B. (2002). Full-information maximum-likelihood estimation of general two-level latent variable models. In preparation.

Bollen, K. A. (1996). An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika, 61,* 109-121.

Joreskog. K.G. & Yang F. (1996). Nonlinear structural equation models: The Kenny-Judd model with interaction effects. In G. Marcoulides & R. Schumaker (eds.), *Advanced Structural Equation Modeling* (pp. 57-87). Mahwah, N.J.: Lawrence Erlbaum Assoc.

Klein, A. & Moosbrugger (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika, 65,* 457-474.

Muthén, B. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modeling. In Collins, L.M. & Sayer, A. (eds.), *New Methods for the Analysis of Change* (pp. 291-322). Washington, D.C.: APA.

Muthén, B. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29,* 81-117.

Muthén, B. & Asparouhov, T. (2002). Modeling of heteroscedastic measurement errors. Mplus Web Note #3.

Muthén, B. & Asparouhov, T. (2003). Integrating multilevel and structural equation modeling. In preparation.

Muthén, L. & Muthén, B. (1998-2002). *Mplus User's Guide.* Los Angeles, CA: Muthén & Muthén.

Muthén, L. & Muthén, B. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 4,* 599-620

Muthén, L. & Muthén, B. (2003). *Mplus Version 2.13: Addendum to the Mplus User's Guide.* Available at www.statmodel.com/support/index.html.

Seltzer, M., Choi, K. & Thum, Y. M. (2002a). Examining relationships between where students start and how rapidly they progress: Implications for conducting analyses that help illuminate the distribution of achievement within schools. CSE Technical Report 560. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing, UCLA.

Seltzer, M., Choi, K. & Thum, Y. M. (2002b). Latent variable modeling in the hierarchical modeling framework: Exploring initial status treatment interactions in longitudinal studies. CSE Technical Report 559. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing, UCLA.

Thum, Y. M. (2002). Measuring Student and School Progress with the California API. CSE Technical Report 578. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing, UCLA.