Modeling of Heteroscedastic Measurement Errors

Bengt Muthén
University of California, Los Angeles

Tihomir Asparouhov
Muthén & Muthén
[*]

Mplus Web Notes: No. 3
Version 1, July 19, 2002

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

1 Introduction

This note presents an approach to handling heteroscedastic measurement errors in latent variable models. The note is motivated by a study of hierarchical modeling of sequential behavioral data, Dagne, Howe, Brown, and Muthén (2002). In Dagne et al. (2002) latent variables are used to represent true log odds ratios for positive and negative responses in subepisodes of couples' behavioral interactions before and after the midpoint of the episode. The precision of the observed log odds ratios as an estimate of the true log odds ratio increases as the number of behaviors in an episode increases. Dagne et al. (2002) give a formula for computing the variance to estimate the corresponding imprecision in measurement. With episode being the unit of observation, this gives rise to the heteroscedasticity of measurement error across episodes.

The approach proposed here draws on the fact that regression with a random slope allows variance changes in an outcome as a function of individual values on the predictor. Maximum-likelihood estimation of such a model can be performed using random slope modeling in Mplus Version 2.1 (Muthén & Muthén, 1998-2002).

2 A Heteroscedastic Measurement Error Model

Consider a latent variable model with heteroscedastic measurement errors for a set of continuous outcomes y_{ij}, where i corresponds to individual and $j = 1, 2, \ldots, p$ varies across the outcomes,

$$y_{ij} = \nu_j + \boldsymbol{\lambda}_j' \, \boldsymbol{\eta}_i + e_{ij} + \epsilon_{ij}, \tag{1}$$

where ν_j is a measurement intercept, $\boldsymbol{\lambda}_j'$ is a row vector of slopes (loadings) for the jth outcome regressed on the m latent variables, $\boldsymbol{\eta}_i$ is the vector of m latent variables, $E(\boldsymbol{\eta}_i) = 0$, $V(\boldsymbol{\eta}_i) = \boldsymbol{\Psi}$, $E(e_{ij}) = 0$, $V(e_{ij}) = \sigma^2_{ij}$, $E(\epsilon_{ij}) = 0$ and $V(\epsilon_{ij}) = \theta_j$. Here, the heteroscedastic measurement error variances σ^2_{ij} vary across individuals i as,

$$\sigma^2_{ij} = \sigma^2_j \; c_{ij}, \tag{2}$$

where σ^2_j is a parameter and c_{ij} are known, fixed values. Normality is assumed for η_i, e_{ij}, and ϵ_{ij}, with the terms being mutually uncorrelated.

3 The Dagne Et Al. Model in the Mplus Random Slopes Framework

Inserting (10) in (8) in Dagne et al. (2002) shows that a special case of (1) - (2) is considered with only two outcomes ($p = 2$) and a single latent variable ($m = 1$) with unit loadings ($\lambda = 1$). Also, σ^2_j is fixed at unity. The relationship between the notation in (1) - (2) and that in Dagne et al. (2002) is: $i = m$, $\nu_j = \mu_j$, $\eta_i = \delta_m$, $\psi = \psi^2$, $e_{ij} = E_{mj}$, $\epsilon_{ij} = V_{mj}$, $\sigma_j^2 = 1$, $c_{ij} = \sigma_{mj}^2$, and $\theta_j = \tau_j^2$. In Dagne et al. (2002), c_{ij} corresponds to the estimated within subepisode standard deviation, where the formula for the variance estimate for σ_{mj}^2 is given in (2) in Dagne et al. (2002). These estimated values are treated as fixed, known values, i.e. their sampling variance is not included in the modeling.

The random slope approach to heteroscedastic measurement errors is as follows for the Dagne et al. (2002) model. For each y_j, rewrite the Dagne et al. (2002) model in line with (1) - (2), but using a random slope regression formulation that fits into the Mplus framework,

$$y_{ij} = \nu_j + \eta_i + \beta_{ij}\ x_{ij} + \epsilon_{ij}, \tag{3}$$

where x_{ij} is a dummy predictor variable containing the known measurement error standard deviations, $x_{ij} = \sqrt{c_{ij}}$, and having a random slope β_{ij} distributed as $N(0, 1)$. This means that the heteroscedasticity is expressed as $V(\beta_{ij}\ x_{ij}|x_{ij}) = V(\beta_{ij})\ x_{ij}^2$, giving the desired density $[\beta_{ij}\ x_{ij}|x_{ij}] = N(0, x_{ij}^2)$. Note that $[y_{ij}|x_{ij}]$ is normal and the mean and variance structure is unchanged compared to (1) - (2),

$$E(y_{ij}|x_{ij}) = \nu_j + E(\eta_i|x_{ij}) + E(\beta_{ij})\ x_{ij} = \nu_j, \tag{4}$$

$$V(y_{ij}|x_{ij}) = V(\eta_i) + x_{ij}^2 + \theta_j. \tag{5}$$

The random slope approach adds a term to the log likelihood corresponding to the added x variable in (3). Log likelihood ratio chi-square testing of nested models is not, however, affected.

As indicated in (1) - (2), it may also be noted that the random slope variance $V(\beta)$ can be estimated providing further measurement error modeling flexibility. Given that the heteroscedastic term $V(\beta_{ij})\ x_{ij}^2 = 0$ for zero β variance, a significant β variance may be taken as evidence of heteroscedasticity. Zero β variance means that (2) is zero and that $e_{ij} = 0$ so that (1) reverts to a conventional latent variable measurement model.

4 Examples

Mplus 2.1 input for the six models estimated in Dagne et al. (2002), and the corresponding data, are available at the Mplus web site, www.statmodel.com.

Model 1 in Dagne et al. (2002) has 5 parameters: ν_1, ν_2, ψ, θ_1, and θ_1. Here, $n = 254$. Freeing the random slope variance for each of the two outcomes, i.e. estimating σ_j^2 in (2), results in an improvement in the log likelihood corresponding to a chi-square difference test (twice the log likelihood difference) of 5.196 with 2 degrees of freedom, which is not significant at the 5% level (the 5% critical value is 5.991). The maximum-likelihood estimates of the random slope variances (standard errors in parenthesis) are 0.861 (.244) and 0.591 (.147). This implies that the Dagne et al. (2002) choice of unit variance values is reasonable, although a bit high for the second outcome. Also, σ_j^2 in (2) appears to be different from zero so that the heteroscedastic measurement model is needed. A conventional homoscedastic measurement model with $\sigma_j^2 = 0$ in (2) and $e_{ij} = 0$ in (1) also uses 5 parameters as in the Dagne et al. (2002) Model 1, but has a lower log likelihood value. The conventional model estimates the reliabilities of the two outcomes as 0.16 and 0.17, respectively. In contrast, Model 1 in Dagne et al. (2002) gives reliabilities that vary across observations. For observations at the minimum, average, and maximum variance estimates of subepisode variance, the reliabilities for the first outcome are 0.27, 0.10, and 0.03. For the second outcome they are 0.28, 0.10, and 0.04.

References

Dagne, G. A., Howe, G. W., Brown, C. H. & Muthén, B. (2002). Hierarchical modeling of sequential behavioral data: An empirical Bayesian approach. Psychological Methods, 7, 262-280.

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