Auxiliary Variables Predicting Missing Data

Tihomir Asparouhov and Bengt Muthén

May 5, 2008

In this document we describe how Mplus includes auxiliary variables in the estimation of structural equation models (SEM) in the presence of missing data. These variables are included in the Mplus analysis using the following command *auxiliary=*$z_1 - z_k$ *(m)*. Examples of such analysis can be found in the V5.1 Addendum to the Mplus User's Guide (Muthen & Muthen 1998-2008) available at statmodel.com.

Suppose that $Y = (Y_1, Y_2, ..., Y_p)$ is the vector of all continuous dependent variables, $\eta = (\eta_1, \eta_2, ..., \eta_m)$ is the vector of all continuous latent variables and $X = (X_1, X_2, ..., X_q)$ is the vector of all covariates. The standard structural equation model is given by the following equations

$$Y = \nu + \Lambda\eta + \varepsilon$$

$$\eta = \alpha + B\eta + \Gamma X + \xi$$

where ε and ξ are zero mean normally distributed residuals with variance covariance Θ and Ψ. Suppose also that there is missing data in the data set that we use to estimate the above SEM model and that there are additional auxiliary variables in the data set $Z = (Z_1, Z_2, ..., Z_k)$ that are not part of the structural model but could be correlated to some of the variables in the model. Because of such potential correlation these variables can be used to reduce the uncertainty caused by the missing data and thereby improve the precision of the estimation. In addition if these variables are related to the missing data mechanism including these variables in the analysis could reduce or eliminate parameter estimates biases that are due to the missing data when the missing data in not missing at random (NMAR). Standard maximum likelihood estimation based only on the variables in the model, i.e., Y and X will lead to unbiased estimates if the missing data is missing completely at random (MCAR) or missing at random (MAR) but will lead to biased estimates if the missing data is NMAR, see Little and Rubin (1987). Facilitating the additional information that auxiliary variables provide can enable us to reduce or eliminate such bias. In addition, to evaluate model fit correctly in the presence of this additional information, the baseline model estimation also incorporates the auxiliary variables, i.e., approximate fit indices such as CFI and TLI can be used in the usual way.

The algorithm implemented in Mplus is based on the maximum-likelihood estimation of Graham (2003) saturated correlates model. We estimate the following expanded model which consists of the two equations of the original model as well as

$$X = \nu_x + \varepsilon_x$$

$$Z = \nu_z + \varepsilon_z$$

where ε_x and ε_z are normally distributed zero mean residual which are also correlated to the ε residuals according to the following variance covariance structure

$$Var(\varepsilon, \varepsilon_x, \varepsilon_z) = \begin{pmatrix} \Theta & 0 & \Theta_{yz} \\ 0 & \Theta_{xx} & \Theta_{xz} \\ \Theta'_{yz} & \Theta'_{xz} & \Theta_{zz} \end{pmatrix} \quad (1)$$

where Θ is the variance covariance in the original model and Θ_{xx} and Θ_{zz} are full unrestricted variance covariance symmetric matrices and Θ_{yz} and Θ_{xz} are also full unrestricted covariance matrices. Estimating this model incorporates the information available in the auxiliary variables without altering the model or without causing any bias in the estimates of the original model.

In the above estimation it is important that the variance covariance matrix (1) is not restricted to a positive definite matrix. This is critical for obtaining unbiased estimates. If the expanded model is misspecified and the auxiliary variables are also correlated to the latent variable η the variance covariance matrix (1) may not be positive definite and if it is restricted to be positive definite the resulting estimates in the original model will be biased. Mplus estimates matrix (1) without any restrictions so that the estimates of the original model are unbiased. The fact that matrix (1) is not positive definite has no effect on the original model nor on the actual model estimated overall variance covariance matrix, which will always be positive definite. See also Savalei and Bentler (2007) for discussion on this topic.

Mplus also uses the auxiliary variables to estimate the unrestricted H1 model which is consequently used for chi-square testing as well as in the computation of the various fit indices. The baseline model used in the computation of the fit indices is also computed using the expanded data set however the actual degrees of freedom are the same as for the original model without the auxiliary variables. In the baseline model the distribution of X and Z is still unrestricted, the Y variables are assumed to be independent among each other and independent from the X variables but their correlation with the Z variables is again unrestricted. The log-likelihood that Mplus computes includes also the auxiliary variables Z and the covariates X and it should not be used for comparison with a model that does not use the auxiliary variables, however it can be used to test against another model that has the same covariates and auxiliary variables. Factor scores are estimated

using the original variables X and Y only, rather than the expanded model which can be misspecified if the latent variables are correlated to the auxiliary variables. The expanded model could have many more parameters than the original model. This could lead to estimation problems of the asymptotic distribution of the parameter estimates in the expanded model. If such a problem occurs and is related to an expanded parameter, that parameter is treated as fixed and the asymptotic distribution of the remaining parameters is computed.

References

Collins, L. M., Schafer, J. L., & Kam, C. (2001). A comparison of inclusive and restrictive strategies in modern missing data. Psychological Methods, 6, 330 - 351.

Enders, C.K., Peugh, J.L. (2004). Using an EM covariance matrix to estimate structural equation models with missing data: Choosing an adjusted sample size to improve the accuracy of inferences. Structural Equation Modeling: A Multidisciplinary Journal, 11, 1-19.

Graham, John W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. Structural Equation Modeling, 10(1), 80-100.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data. New York: John Wiley.

Muthen, L.K. & Muthen, B.O. (1998-2008). Mplus User's Guide. Fifth Edition. Los Angeles, CA: Muthen & Muthen

Savalei, V., and Bentler, P. (2007). A Two-Stage ML Approach to Missing Data: Theory and Application to Auxiliary Variables. Department of Statistics Papers, UCLA.