Using Mplus Monte Carlo Simulations In Practice: A Note On Assessing Estimation Quality and Power In Latent Variable Models

Bengt Muthén
University of California, Los Angeles
*

Mplus Web Notes: No. 1
Version 2, March 22, 2002

*The research was supported under grant K02 AA 00230-01 from NIAAA. The email address is bmuthen@ucla.edu.

1 Introduction

Researchers often ask for rules of thumb to guide them in deciding if parameter estimates should be considered large enough to be important, if standard errors can be trusted with small samples or non-normal data, and in deciding if the sample is large enough to detect a hypothesized effect. Rules of thumb, however, are often not targeted to the research problem at hand. This note shows that Monte Carlo simulations are useful for not only methodologists but also for substantive researchers, making it possible to answer such questions in a convincing way.

In Monte Carlo studies data are generated from a population with hypothesized parameter values. A large number of samples are drawn, and for each sample a model is estimated. This makes it possible to assess the quality of parameter estimates, standard errors, tests of model fit, and power. A substantive research study can benefit from augmenting the data analysis by a Monte Carlo study in order to evaluate the findings. In this case, the parameter values estimated from the data may serve as the best guess of the population parameter values. Using Monte Carlo simulations, a researcher can answer questions about which sample size would have been needed, or would have been sufficient, to have acceptable quality of estimates and power to reject zero effects. By tailoring the Monte Carlo study to the specific case at hand, the researcher avoids having to rely on general rules of thumb that may match their situation poorly.

The quality of parameter estimates and standard errors may deteriorate as a function of a variety of factors, including small sample size and missing data. This deterioration influences the assessment of significance of coefficients, often finding significance where it is not present. Power is also influenced by these data-related factors, and is in addition influenced by factors related to the model such as parameter values and number of observed variables or number of time points in longitudinal studies. Power computations in growth model contexts were presented in Muthén and Curran (1997) using the Satorra-Saris method. A description of how to use this method is given on the Mplus web site at www.statmodel.com/power.html. This method does not require a Monte Carlo study. Because it computes power from a model's population means, variances, and covariances, however, the Satorra-Saris method is not suitable when the analysis calls for information beyond the second-order moments, as in analysis with missing data and non-normal data.

This note describes some of the Mplus Monte Carlo facilities. Using mixture analysis, a very flexible Monte Carlo procedure is available. This includes conventional latent variable modeling with a single class as a special case, which is the topic to be studied here. A powerful feature in mixture Monte Carlo is that data can be generated by a certain model, missing data added according to a second model, and the resulting data analyzed according to a third model. As an illustration, a simple linear growth model using maximum-likelihood estimation under normality assumptions will be studied. The Monte Carlo analyses show how estimation quality and power are influenced by varying sample size and missing data. Model tests of fit will not be studied in this note. While a growth model is used for illustrative purposes here, the approach can be carried out using factor analysis models, MIMIC models, structural equation models, latent class analysis, growth mixture analysis, etc.

Three Monte Carlo studies of the linear growth model are carried out using maximum-likelihood estimation under normality assumptions. Mplus input and output files referred to in the text are found at www.statmodel.com, under Mplus Web Notes. Mplus input and output will be explained, and conclusions drawn related to estimation quality and power. The input setups can then be easily generalized to other situations. All examples can be run using the free Mplus Demo version available at

www.statmodel.com/mplus/demo.html.

A summary of the Mplus language can be obtained at

www.statmodel.com/mplus/language.html.

2 Monte Carlo Results

Following is a description of the results of a Monte Carlo study in the Mplus mixture track (for a technical description of Mplus mixture Monte Carlo, see Muthén & Muthén, 1998-2001; Appendix). First, the average of parameter estimates over the repeated draws of independent samples, referred to as replications, is reported for each parameter. The average can be compared to the true value that was used in generating the sample in order to study bias. Second, the standard deviation of parameter estimates over the repeated draws of independent samples is reported for each parameter. With a sufficiently large number of replications, the standard deviation of estimates may be seen as the true value for the variability across replications of the parameter estimates. Third, the average over replications of the estimated standard error of the parameter estimate is reported for each parameter. The average standard error estimate can be compared with the standard deviation of estimates to check the bias in the estimated standard errors. Fourth, the mean square error of estimates is reported as a summary of both bias and variability, where

$$\sum_{r=1}^{R}(\hat{\pi}-\bar{\hat{\pi}})^2/R+(\bar{\hat{\pi}}-\pi)^2, \tag{1}$$

where π is the parameter in question and R is the number of replications. Fifth, the 95% coverage is reported, i.e. the proportion of the replications where a 95% confidence interval covers the true parameter value. Sixth, the significance at the 5% level (two-sided test) is considered for each estimate,

$$z=\frac{\hat{\pi}-0}{s.e.(\hat{\pi})}, \tag{2}$$

where the 0 value is included to emphasize that the significance test in this part of the output tests against a 0 value. Assuming that z is approximately normal, a value is deemed significant if the absolute value of z exceeds 1.96. The proportion of significant values across the replications is reported. The proportion of significance is the Monte Carlo determined power to reject the hypothesis that $\pi = 0$. Power values of at least 0.80 are generally viewed as desirable.

Results differ in the details depending on the choice of seed value for the data generation and the choice of the number of replications. The seed influences the random samples drawn from the population, producing different starting points for the random draws. Any randomly chosen seed can be used. Regarding the number of replications, a conservative choice is $10,000$ replications, although frequently 500 replications is sufficient and will speed up the computations. Experimentation with the model in Study A indicated that $1,000$ replications, although sufficient to study parameter estimates and standard errors, was not quite large enough to produce power values with two dependable decimals. With $10,000$ replications, three different seed values gave the same power values to two decimals. A useful strategy in practice is to first try a single replication to check that parameter estimates are reasonable, catching possible errors in the model set up. Next, 100 replications can be tried, providing a way to estimate the time required to do a larger number of replications.

3 Study A. Linear Growth Model: No Covariates, No Missing Data

The Monte Carlo simulation labelled mca (see web site) concerns a linear growth model with 4 time points, no covariates, and no missing data. It is of particular interest to study the mean of the slope factor with respect to estimation quality and power. The following is an explanation of the model choice, the Mplus input, and the resulting output. Maximum-likelihood estimation under normality is used throughout all three studies.

3.1 Choice of model

The model specified in the analysis is a linear latent variable growth model with positive growth over 4 time points with random intercept and slope factors (for Mplus-related growth model literature, see /www.statmodel.com/references.html#growth and www.statmodel.com/references.html#growthmixture). The outcomes are normally distributed. The time points are equidistant. The outcome intercepts are fixed at zero. The true growth factor means are 0 and 0.2, respectively. The growth factor variances are 0.5 and 0.1, respectively, reflecting a common variance ratio. The growth factors have zero covariance. The outcome residual variances are all 0.5. This gives a time 1 R^2 in the outcome of 50%, increasing over time due to increasing growth curve variance. The standard deviation of the outcome at time 1 is 1.00 and the mean increase over the four time points is 0.60.

3.2 Mplus input

The TITLE command provides a title for the output.

The MONTECARLO command describes the Monte Carlo study. The NAMES option assigns names to the variables in the generated data. The NOBS command specifies the number of observation in the random samples. The NREPS option specifies the number of replications in the analysis. The SEED option selects the seed to be used for random sampling. The NCLASSES option specifies the number of classes in the analysis model. For conventional growth modeling, there is one class. The GCLASSES option specifies the number of classes in the data generation model. The SAVE option specifies a file where the data from the first replication is stored.

The ANALYSIS command provides information about the type of analysis to be performed. By selecting mixture, a mixture model will be estimated. With a single class, this is a conventional latent variable model. The mixture model automatically includes a mean structure for the y variables. The estimator option specifies that regular maximum-likelihood estimation is to be performed, resulting in regular standard error estimation under multivariate normality.

The MODEL MONTECARLO command is used to provide the true population parameter values to be used in the data generation. The same options are used as in the MODEL command. Each population parameter must be specified followed by the @ symbol or the asterisk and the true parameter value. The remaining model specifications follow that of a regular growth model (see Muthén & Muthén, 1998-2001). The two BY statements define the intercept and slope factors in the growth model. The loadings of the intercept factor are fixed at one. The loadings of the slope factor are fixed at 0, 1, 2, and 3 to define a linear growth model with equidistant time points. The default zero loading for the slope factor at time point one defines the intercept factor as an initial status factor. Variables in square brackets refer to the means or intercepts of those variables, while variables without square brackets refer to the variances or residual variances of the variables. The intercepts of the outcome variables are fixed to zero, capturing the mean change over time via the growth factor means. The residual variances of the outcome variables are different across time, and the residuals are not correlated as the default. Note that as a function of using the Mplus mixture track, the growth factor means need to also be respecified within the first (and only) class to avoid the default setting of zero factor means in the first class.

The MODEL command describes the model to be estimated. In this situation, the data generation model and the analysis models are the same so that the analysis model is correctly specified for the data generated. The asterisks denote the parameters to be estimated, i.e. the two growth factor means, the two growth factor variances, the growth factor covariance, and the four residual variances for the outcomes. The starting values given for these parameters are the true values given in the MODEL MONTECARLO command. Coverage values will be computed based on these starting values.

The OUTPUT command is used to obtain additional output beyond the Monte Carlo summaries over the replications. The TECH9 option produces error messages related to convergence for each replication of the Monte Carlo study.

3.3 Mplus output

The output shows that all of the requested replications were completed successfully. The sample statistics output refers to the first replication and shows the expected increasing mean trend and increasing variances. The log likelihood value is also for the first replication.

The MODEL RESULTS section of the output gives the summaries over the replications. Here, the column labelled Starting gives the true parameter values that generated the data. The remaining six columns were described above in section 2, Monte Carlo Results.

The results indicate negligible parameter estimate bias. The standard errors show a small downward bias that seems ignorable for most practical purposes (typically on the order of 5%). This bias disappears as the sample size increases sufficiently (this can be experimented with by changing NOBS = 50 to say NOBS = 1000). The coverage ranges from 92% to 94% over the parameters, which seems quite acceptable. The last column gives the power values. For example, it is seen that the power to reject a zero mean for the slope factor reaches an acceptable value of 0.89 at this sample size. This value may be compared to that obtained by the Satorra-Saris method given at www.statmodel.com/power.html, where the same model and parameter values are considered. The Satorra-Saris power value for $n = 50$ is 0.85. As pointed out in Muthén and Curran (1997, pp. 382-383), the Satorra-Saris power estimates are less trustworthy for small sample sizes, say below 100. It may be noted, however, that the value of 0.89 from the Monte Carlo study may be slightly inflated due to the reported slight underestimation of the estimated standard error (in fact, if an ad hoc adjustment of the 0.89 value is done based on the standard error bias factor of 0.0615/0.0641, the value 0.85 is obtained).

Based on this setup, it is easy to study several variations on the theme. Changing the sample size from $n = 50$ to $n = 25$ gives a Monte Carlo power estimate of 0.64, which (even if somewhat inflated) is not generally deemed sufficient. Changing the population mean of the slope factor from 0.2 to 0.1, corresponding to an increase over the four time points of 0.3 time 1 standard deviation units, gives a power estimate of 0.38 at $n = 50$ and 0.89 at $n = 200$. A researcher may also want to study the effects of using more time points, either extending the time span of the study or adding more frequent measurement occasions.

4 Study B. Linear Growth Model: Covariate, No Missing Data

The Monte Carlo simulation labelled mcb (see web site) represents minor variations on the theme of Study A. Here, a covariate is added to the model and it is of interest to study the influence of the covariate on the slope factor. Only the new parts of the model, the input, and the output will be described.

4.1 Model choice

The Study B model adds a time-invariant covariate x to the linear growth model of Study A. The growth factor means and variances are kept at the same values. The R^2 value for the intercept factor is 50% while it is 10% for the slope factor, reflecting values commonly seen in practice. The x variable is normally distributed with mean zero and variance one. A sample size of $n = 200$ is considered.

4.2 Mplus input

Following are changes due to including the covariate x.

The NAMES are option now includes x.

The MODEL MONTECARLO command specifies the mean and the variance of the independent variable x in the overall model.

In the MODEL command, note that the mean and variance of the x variable should not be referred to in the analysis model because the marginal distribution of the x variable is not part of the analysis model.

4.3 Mplus output

The MODEL RESULTS section of the output shows that the parameter estimate and standard error biases are negligible. The coverage is between 94% and 95% in all cases. The estimate of the power for the regression of the slope factor on the covariate is 0.90 (0.898), reflecting a sufficient power to reject a hypothesis of zero influence for this sample size. A sample size of $n = 100$ gives a power estimate of 0.64.

5 Study C. Linear Growth Model: Covariate, MAR Missing Data

The Monte Carlo simulation labelled mcc (see web site) represents minor variations on the theme of Study B. Here, the effect of missing data on the outcomes is studied. Only the differences in the model, the input, and the output will be described.

5.1 Model choice

Study C adds missing data to Study B. The missingness appears among the outcomes y. For y_1 missingness is MCAR, while for $y_2 - y_4$ the missingness is MAR and predicted by x (for MCAR and MAR definitions, see Little & Rubin, 1987). The choice of missingness model reflects the increasing missingness and selective dropping out often seen in longitudinal studies. The influence of the covariate on the missingness is described by logistic regression, where the probability of missingness (P) is expressed as

$$P = 1/(1 + e^{-L}), \tag{3}$$

where $L = \alpha + \beta\, x$ is the logit and α and β are the logit regression intercept and slope, respectively. In Study C, the logit slopes are set at 1 for the last three time points. The slope is zero at time 1. At time 1, the choice of $\alpha = -2$ gives an expected 12% missing data. Considering as an example missingness at time 4 for which $\alpha = 0$, this means that for a one standard deviation covariate increase from the mean of zero the percent missing at time 4 increases from 50% (logit value 0) to 73% (logit value 1). At the covariate mean of zero, the logit values of Study C produce an expected missingness of 18% at time 2, 27% at time 3, and 50% at time 4. It is possible to include more than one missingness covariate and the missingness covariates need not be the same as the data generation or analysis covariates.

5.2 Mplus input

Following are the changes due to including the missingness.

The MONTECARLO command includes the option missing = y1 - y4.

The ANALYSIS command specifies the option type = mixture missing.

The MODEL MISSING command specifies the probability of missing data on the outcomes given in logit scale. For y_1, the bracket logit value of -2 translates into a probability of 0.12 using the formula $P = 1/(1 + e^{-L})$, where P is the probability and L is the logit. For $y_2 - y_4$ four different logistic regressions describe the probability of missingness as a function of x (the same formula can be used for translation into probabilities). The bracket value gives the intercept and the ON statement gives the slope of the logistic regression.

5.3 Mplus output

The output now contains information from the first replication about missing data patterns as well as sample coverage for each variable and pairs of variables. The parameter estimates bias is negligible, while there is a slight underestimation seen in the standard errors. The coverage values are good, ranging from 93% to 94%. The power value for the slope factor regression on the covariate is estimated at 0.75, which may be compared to the complete-data estimate of 0.90. Larger sample sizes can be experimented with to find the sample size required to attain the desirable power level of 0.80.

6 Discussion

This note provides a description of straightforward ways to study estimate quality and power using Monte Carlo simulations in Mplus. Many other variations on the three studies are possible. A covariate can be dichotomized to illustrate the effects of a dummy covariate such as gender or treatment condition. The analysis model may be misspecified as compared to the data generation model to study biases in estimates. Missing data patterns corresponding to multiple-cohort settings can be generated (for an example, see the Mplus User's Guide, Example 29.2A, p. 325) in order to study power of alternative designs. Non-ignorable missing data can be generated as a function of latent class variables to study biases when using the standard MAR assumption. These options are described in the Mplus User's Guide. In some cases, a researcher may want to generate data in ways not covered by Mplus. Monte Carlo studies with data generated outside Mplus are aided by the Mplus RUNALL utility described at www.statmodel.com/runutil.html. Hopefully, this note can inspire both substantive and methodological researchers to interesting Monte Carlo studies.

References

Little, R.J., & Rubin, D.B. (1987). Statistical analysis with missing data. New York: John Wiley & Sons.

Muthén, B. & Curran, P. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. Psychological Methods, 2, 371-402.

Muthén, L. & Muthén, B. (1998-2001). Mplus User's Guide. Los Angeles, CA: Muthén & Muthén.