Assessing model fit for SEM models with categorical variables via contingency tables

Tihomir Asparouhov & Bengt Muthén

April 26, 2022

1 Introduction

Test of fit for models with categorical variables can be obtained in two ways. The traditional way, is based on evaluating the statistical significance of the difference between $Var(Y^*|H1)$ and $Var(Y^*|H0)$. Here Y^* is the underlying continuous variable that is cut to obtain the categorical values. The existence of such variable is guaranteed by the probit link function. The $H0$ model is the SEM model and the H1 model is the multivariate unrestricted probit model. The H1 model can be estimated in Mplus with the WLS type estimators as well as the Bayes estimator. For the WLS type estimator we obtain a chi-square value while in the Bayesian case we obtain a PPP value.

An alternative way to evaluate the model fit is to consider the differences between the two distributions of the observed values Y instead of the latent variable Y^*. The distribution of Y is essentially the contingency (multivariate frequencies) tables. Therefore, model fit can be established by comparing the observed and the estimated contingency tables. Such comparisons are available in Mplus for all three of the estimators available for estimating SEM with categorical variables: ML, WLS(MV), and Bayes. The contingency table comparisons are obtained with the option OUTPUT:TECH10. There are three types of contingency tables reported in that output: univariate tables, bivariate tables, and the full multivariate table.

There are three ways these results can be used. First, for every cell in every contingency table, Mplus computes a standardized residual estimate (Z-score). Thus, one can determine which frequency cells are not fitted sufficiently well and modify the model accordingly. Here one should be aware of

the fact that the tables are subject to multiple testing issues. Even though the cell tests are not completely independent of each other, we can generally expect that one out of every 20 cell will be misfitted at the nominal 95% level.

The contingency tables can also be used for comparative purposes. One can compare for example, across different models and even estimators, the total number of misfitted cells (standardized Z-score above 1.96) or the total sum of all Pearson chi-square statistics across all (univariate and bivariate) contingency tables, see equation (44) in Asparouhov and Muthén (2021). Although such a comparison does not have an associated statistical significance, it is nevertheless a valuable way to quantify the fit of the model. This approach is unique in being able to compare models across different estimators. Both the total number of misfitted univariate and bivariate cells as well as the total Pearson chi-square statistic for the univariate and bivariate tables are reported in the Mplus output. The full contingency table is somewhat more difficult to use. It is quite common to have many of the cells in that table have very low frequency, simply because the full table will have a very large number of cells. To make this somewhat easier to use, Mplus also produces a table with 20 most frequent patterns from the full contingency table as well as the total number of cells misfitted in those 20 patterns (Z-score above 1.96), which can also be used for comparative purposes.

The third way the contingency table results can be used is via proper statistical tests. Such tests are available in two cases. The first case is the classic Pearson/Chi-Square test of fit for the full contingency table available in Mplus with the ML estimators. The second case is the Pearson PPP described in Section 3.8 in Asparouhov and Muthén (2021) which uses as the fit statistic the sum of univariate and bivariate Pearson statistics for all univariate and bivariate contingency tables.

One important question that arises is how the classic latent Y^* testing and the tech10 observed Y testing compare. In many situations the two will be comparable and will lead to the same conclusion, however, there are situations where the two versions are completely different. We will mention just two situations here to illustrate the possibilities but these are by no means the only reasons the two statistics will differ. The first situation is the case where the Y^* testing does not reject the model while the Y testing rejects the model. This can happen when the H1 multivariate probit model is not a sufficiently good fit for the observed data. This happens when the data is for example generated from a Mixture model, and is analyzed as a single-class

model. It will also happen when the data is generated with two-part ordinal model but is analyzed with standard SEM model based on fitting the data via the probit link function. The second situation is where the Y^* testing rejects the model while the Y testing does not reject the model. An example of this is a SEM model that includes both continuous and categorical indicators. The Y contingency table testing only concerns the categorical variables and that part of the model can be fitted well, while the Y^* testing includes the continuous variables which can be misfitted. Here the role of the continuous dependent variable can also be taken by covariates. The contingency tables do not provide information on how well the covariates predict the outcomes. Thus, it would not be unusual to have well fitted contingency tables that are rejected by the classic chi-square due to problems with the predictive part of the model.

In this paper we provide several technical details on the computations implemented in Mplus. First, we focus on the computation of the estimated contingency tables. Depending on the estimator, different methods are used. Second we show how the standardized residuals are computed and interpreted. Finally we describe the computation for two special models, the two-part ordinal model and the model with observed mediators.

2 Estimated contingency tables

The computations vary somewhat between the different estimators. First we consider computation with the Bayes estimator using probit link functions. Let's first consider the univariate contingency tables. The estimated SEM model can generally be represented as

$$Y^* = \alpha + \beta Z + \varepsilon$$

where ε represents the residual and Z represents all predictors for Y including latent predictors, other observed dependent variable predictors and covariates. From here we compute the unconditional distribution of $Y^* \sim N(\mu, \sigma)$ assuming normal distribution for Z and ε. At this point the estimated probability is given by

$$P(Y = i) = \Phi((\tau_i - \mu)/\sqrt{\sigma}) - \Phi((\tau_{i-1} - \mu)/\sqrt{\sigma}),$$

where τ are the threshold parameters and Φ is the standard normal distribution function.

For the bivariate probability we similarly derive from the model $(Y_1^*, Y_2^*) \sim N(\mu, \Sigma)$ from the structural model where μ is a vector of size 2 and Σ is a 2 by 2 variance covariance matrix.

$$P(Y_1 = i_1, Y_2 = i_2) = \Phi_{2,\rho}((\tau_{i_1,1} - \mu_1)/\sqrt{\sigma_{1,1}}, (\tau_{i_2,2} - \mu_2)/\sqrt{\sigma_{2,2}})) +$$

$$\Phi_{2,\rho}((\tau_{i_1-1,1} - \mu_1)/\sqrt{\sigma_{1,1}}, (\tau_{i_2-1,2} - \mu_2)/\sqrt{\sigma_{2,2}})) -$$
$$\Phi_{2,\rho}((\tau_{i_1,1} - \mu_1)/\sqrt{\sigma_{1,1}}, (\tau_{i_2-1,2} - \mu_2)/\sqrt{\sigma_{2,2}})) -$$
$$\Phi_{2,\rho}((\tau_{i_1-1,1} - \mu_1)/\sqrt{\sigma_{1,1}}, (\tau_{i_2,2} - \mu_2)/\sqrt{\sigma_{2,2}})),$$

where $\Phi_{2,\rho}$ is the bivariate distribution function for a normal distribution with variances of 1 and correlation ρ. With logit link function and the Bayes estimator, the residual probit variance of 1 is replaced by the logit variance of $\pi^2/3$. There are some other variations in this computation across the different estimators. With the WLSMV estimators, the covariates are not integrated with the rest of the variables but are conditioned on. This means that the probabilities are computed for each individual conditional on the covariates and are averaged across the sample. Such a computation is somewhat more precise but also would be computationally more intensive. With the ML estimator we express the SEM model slightly differently. For the univariate probability the model is expressed as

$$P(Y = i|Z) = \Phi((\tau_i - Z_0)) - \Phi((\tau_{i-1} - Z_0),$$

where $Z_0 = \alpha + \beta Z \sim N(\mu, V)$. Then the probability is computed as

$$P(Y = i) = \int \Phi((\tau_i - \mu - Z\sqrt{V}))\phi(Z)d(Z) - \int \Phi((\tau_{i-1} - \mu - Z\sqrt{V}))\phi(Z)d(Z),$$

where ϕ is the standard normal density function. The two integrals are then computed numerically with a 1-dimensional numerical integration. Similar variation is used for the bivariate probability which leads to a 2-dimensional numerical integration.

The estimated univariate and bivariate tables are compared to their observed quantities. If there is missing data, however, the observed quantities might not be comparable to the estimated. The univariate observed tables are computed with listwise deletion. This means that missing values are simply deleted. In the bivariate case the quantities in the observed table are computed from those observations where both variables are observed.

It is well understood that such listwise approach may lead inaccurate values if the missing data is not MCAR. Therefore the comparisons between the estimated and observed univariate and bivariate tables is primarily valid when the missing data is MCAR or a small deviation of MCAR or when the amount of missing data is small.

The full patterns estimated frequencies are computed as follows. With the Bayes and WLSMV estimators the model estimated Y^* joint distribution is derived from the estimated structural model for the full vector of underlying latent variables $Y^* \sim N(\mu, \Sigma)$. Here the dimensions of μ and Σ is the number of categorical variables in the model p. A random sample with 10000 draws is taken from this multivariate normal distribution which are cut according to the estimated thresholds to obtain a sample of categorical data based on the estimated model. The frequencies of this sample are then computed to produce the model estimated frequencies. With the ML estimator a different approach is used. Conditional on the latent variables and other predictors, in the ML framework, the categorical variables are independent of each other because WITH statements are not allowed in that framework. In that case the conditional probability is just the product of the individual probabilities

$$P(Y_1 = i_1, Y_2 = i_2, ..., Y_p = i_p|*) = P(Y_1 = i_1|*)P(Y_2 = i_2|*)...P(Y_p = i_p|*).$$

The individual probabilities are directly computable from the model. To get the unconditional probability the above equation is integrated out with the integration method used for the model estimation, i.e., if there is one latent variable, a one dimensional numerical integration is used, etc.

In the presence of missing data, the full contingency table produced by Mplus will contain also the estimated probability for the particular missing data pattern in addition to the actual observed values. It is important to note here that the estimated model does not condition on the missing data pattern. It simply computes the probability of the observed values. If the missing data is not MCAR but it is MAR and the missing data pattern was known a different quantity would be produced. None of the Mplus estimation methods however produce a missing data mechanism estimation and therefore such computation that is conditional on the missing data pattern is not possible. Therefore, we can say here that the estimated probability is computed as if the missing data mechanism in MCAR. Furthermore, since the actual quantities that are compared is the Mplus output are not the probabilities but the frequency counts, the issue arises then regarding how to convert

the probability to a frequency count. To do that, one needs an estimate for the number of observations in the missing data pattern. For this purpose we use the quantity found in the sample. We explicate below with an example. Suppose that there are 3 categorical variables in the model and we need to estimate the frequency of the cell $[Y_1 = i_1, Y_2 = missing, Y_3 = missing]$. The frequency count is estimated as follows

$$P(Y_1 = i_1)P(Y_1 = observed, Y_2 = missing, Y_3 = missing)N = P(Y_1 = i_1)N_1,$$

where N is the total sample size and N_1 is the number of observations in the sample that have the pattern $[Y_1 = observed, Y_2 = missing, Y_3 = missing]$. This kind of comparison is valid when the missing data is MCAR or when the deviation from MCAR is small or when the amount of missing data is small. When the missing data is substantially different from MCAR and the amount of missing data is substantial as well, the comparison is not valid and the differences in the frequency counts maybe due to the missing data mechanism.

There are some extreme examples that are worth mentioning here. If the sample contains a single observation of a particular missing data pattern, such as for example $[Y_1 = observed, Y_2 = missing, Y_3 = missing]$, the estimated quantity is $P(Y_1 = i_1)$, while the observed quantity is 1. Clearly such a comparison is not a reliable indicator of model fit.

3 Z-score for cells in the contingency tables

For each contingency table cell we can easily compute the probability that the observed values occurs given the estimated probability. Suppose that the cell probability is q. We form the binary indicators W_i, where W_i is 0 if the $i-$th observation does not belong in the cell and it is 1 if the $i-$th observation belongs in that cell. The observed probability is then

$$W = \frac{W_1 + W_2 + ... + W_N}{N}.$$

Using the law of large numbers, the distribution of W is approximately $N(q, q(1-q)/N)$ and the standardized Z-score is computed as

$$\frac{W - q}{\sqrt{q(1-q)/N}}.$$

This approach is used in all contingency tables: univariate, bivariate and full pattern. The formula is also applied in the situation where we compare frequency counts rather than probabilities (which are essentially multiples of the probabilities).

The full contingency table has to also deal with two other issues. When the number of variables is large the contingency table is very large and a large number of cells have very small estimated probabilities. These are rare outcomes. The variance of W for such rare outcomes is difficult to estimate. We use an alternative approach to the above formula

$$Var(W) \approx max(q_o(1-q_o)/N + q(1-q)/10000, q(1-q)/N + q(1-q)/10000).$$

In the above formula we have added the variance for q because that is also estimated from the montecarlo draws and there is uncertainty in that estimate. Also, we use the observed frequency q_0 as an alternative estimate for the variance estimation purposes to prevent divisions by zero when the montecarlo draws do not contain the rare outcome.

4 The two-part ordinal model

The two-part ordinal model is a special case that requires a different computation for the purposes of contingency table comparison. This computation is implemented with the Bayes estimator. Consider a two part ordinal variable Y. The variable is decomposed as two categorical variables Y_0 and Y_p. If $Y = 0$ then $Y_0 = 0$ and Y_p is missing. If $Y > 0$ then $Y_0 = 1$ and $Y_p = Y$. If Y is missing, both Y_0 and Y_p are missing. Essentially, the categorical variable Y is replaced by two categorical variables Y_0 and Y_p and the first category of Y (namely the zero category) is treated as a special category that deserves its own categorical variable. Multivariate models with large amount of 0 observations often can not be fitted well by the multivariate probit model and thus the two-part model offers a more flexible framework that nevertheless is much more parsimonious that a contingency table model. In the multivariate probit model the variable Y is represented by a single normal variable Y^*, while in the two part ordinal model the variable is represented by two normally distributed latent variables Y_0^* and Y_p^*. In the multivariate two-part model, more correlations can be added to the model between the underlying latent variables, making it more flexible than the multivariate probit model. If Y and Z are two-part ordinal variables,

the bivariate two-part model can estimate these 4 correlations $Corr(Y_0^*, Z_0^*)$, $Corr(Y_p^*, Z_p^*)$, $Corr(Y_0^*, Z_p^*)$, $Corr(Y_p^*, Z_0^*)$. The correlations $Corr(Y_0^*, Y_p^*)$ and $Corr(Z_0^*, Z_p^*)$ are generally unidentified because when Y_p is observed Y_0 is constant, and the same applies to Z. The bivariate probit model offers just one correlation $Corr(Y^*, Z^*)$.

In principle, it is possible to consider the contingency table fit for Y_0 and Y_p, however, there are two reasons why this is not a good idea. First, Y_p has a very strong MAR missing data and that essentially nullifies the contingency table comparison (which ignores the known missing data mechanism). Second, the contingency tables for Y_0 and Y_p are less desirable as they do not directly address the fit to the data. Only contingency tables for Y can do that.

The estimation of the univariate, bivariate and full pattern tables are estimated as described earlier for the full pattern estimates. That is, the multivariate distribution for $(Y_0^*, Y_p^*) \sim N(\mu, \Sigma)$ is derived from the model, a random sample is drawn from that distribution with 10000 observations, the estimated thresholds are used to cut the variable to produce the observed values for Y_0 and Y_p, the observed values are then converted to the observed value of Y according to the two-part conversion, and finally the population of 10000 Y values is used to obtain estimated univariate, bivariate, and full contingency tables.

For applications of two-part ordinal modeling, see Mplus Web Talk No. 4, Part 2, `https://www.statmodel.com/Webtalk4P2.shtml`.

5 Observed mediator models

The observed mediator models available in Mplus with the option mediator=observed, also require a special technique for estimating the contingency tables. With this modeling option, when a categorical variable Y is used as a predictor, the actual observed value Y is used and not the latent underlying variable Y^* (which is what Mplus will use by default with the Bayes and WLSMV estimators). Here again we obtain the estimated contingency tables using a montecarlo method. The first step in the computation is to order the categorical variables in precedence so that Y_p can be predicted by Y_1,..., Y_{p-1}, Y_{p-1} can be predicted by Y_1,..., Y_{p-2}, etc. Then, using the estimated model, we obtain the distributions for Y_1^*, $[Y_2^*|Y_1]$, $[Y_3^*|Y_1, Y_2]$, etc. All of these distributions are conditional univariate normal distribution. We then

generate a sample with 10000 observations where the categorical variables are generated sequentially from Y_1,..., Y_p, one at a time, from the corresponding conditional distributions. The generated 10000 observed vectors are then used to construct the estimated contingency tables.

The contingency table fit can be used to compare the observed and latent mediator models. Such examples are given in Mplus Web Talk No. 4, Part 2, `https://www.statmodel.com/Webtalk4P2.shtml`.

References

[1] Asparouhov, T., & Muthén, B. (2021). Residual structural equation models.
https://www.statmodel.com/download/Asparouhov_Muthen_2021a.pdf