Chi-Square Statistics with Multiple Imputation

Tihomir Asparouhov and Bengt Muthén

Version 2

July 27, 2010

1 Likelihood Ratio Test

In this section we describe how Mplus computes the likelihood ratio test (LRT) with multiple imputations. The LRT is computed only for the ML estimator for single level SEM models using the method described in Meng and Rubin (1992), see also Enders (2008). For these models the LRT is computed for the estimated model against the unrestricted mean and variance/covariance model, i.e., the usual test of fit. For all other estimators and models the fit statistic is not computed but the distribution of the test statistic over the different imputed data sets is reported.

The LRT test statistics is computed as follows. Suppose that there are M imputed data sets. Let T_m be the test of fit statistic for the $m-$th imputed data set. Let the parameter estimates of the H_0 and H_1 models, using the $m-$th imputed data set, be Q_{0m} and Q_{1m}. Let the number of parameters for the H_0 and H_1 models respectively be p_0 and p_1. Define the average quantities as

$$\bar{T} = \frac{1}{M} \sum_{m=1}^{M} T_m$$

$$\bar{Q_0} = \frac{1}{M} \sum_{m=1}^{M} Q_{0m}$$

$$\bar{Q_1} = \frac{1}{M} \sum_{m=1}^{M} Q_{1m}$$

Now compute the LRT test statistic for the H_0 model against the H_1 model where the parameter estimates are fixed to $\bar{Q_0}$ and $\bar{Q_1}$ respectively, using the $m-$th imputed data set and denote this test statistic value by T'_m. This statistic is averaged over all imputed data sets

$$\bar{T}' = \frac{1}{M}\sum_{m=1}^{M} T'_m.$$

Then final test statistics is

$$T_{imp} = \frac{\bar{T}'}{1+r_3} \tag{1}$$

which has approximately a chi-square distribution with the same degrees of freedom as the usual test of fit statistics, i.e., $p_1 - p_0$. The correction factor r_3 is computed as follows

$$r_3 = \frac{M+1}{(M-1)(p_1-p_0)}(\bar{T} - \bar{T}').$$

The above approximation may be quite poor if the amount of missing data is relatively large or the number of imputations M is low.

2 Wald Test

For imputed data the Wald test is computed for all estimators and models. The computation is based on estimating the joint distribution of the parameter estimates. Suppose that we need to test $F(Q) = 0$ where F is a multivariate function and Q are the model parameters. Let the parameter estimates for the $m-$th imputed data set be Q_m and their estimated asymptotic distribution be V_m. The combined estimates are computed as

$$\bar{Q} = \frac{1}{M}\sum_{m=1}^{M} Q_m$$

and their asymptotic distribution is computed as follows

$$V = \frac{1}{M}\sum_{m=1}^{M} V_m + \frac{M+1}{M(M-1)}\sum_{m=1}^{M}(Q_m - \bar{Q})(Q_m - \bar{Q})^T.$$

The correct Wald test can now be computed the usual way

$$W = F(\bar{Q})(F'(\bar{Q})V(F'(\bar{Q}))^T)^{-1}F(\bar{Q})^T$$

where F' is the first derivative of F with respect to the parameters Q. Under the null hypothesis, $F(Q) = 0$, the distribution of W is a chi-square with d degrees of freedom, where d is the dimension of the restrictions F. In the above formula the delta method is used to obtain the asymptotic covariance of $F(Q)$ from the asymptotic covariance of Q. This method however does not utilize the actual chi-square values across the different imputations as in the previous section. Instead, it computes the Wald test simply by using the estimated asymptotic variance of the parameters, just as the Wald test is computed for complete data analysis.

3 A simulation study

In this section we conduct a simulation study to evaluate the performance of the imputation LRT statistic T_{imp} given in (1). As a comparison we also use the $\bar{T}$ statistic, which is simply the averaged of the chi-square statistics across the imputed data sets. In the simulation study we use the following two factor analysis model. Let the two factors in the model be η_1 and η_2 and let each factor has three observed indicator variables. The model is given by the following two equations. For $j = 1, ..., 3$

$$y_j = \mu_j + \lambda_j \eta_1 + \varepsilon_j \tag{2}$$

and for $j = 4, ..., 6$

$$y_j = \mu_j + \lambda_j \eta_2 + \varepsilon_j. \tag{3}$$

where y_j are the observed variables, ε_j are the residual variables, μ_j are the intercept parameters and λ_j are the loading parameters. We generate 100 data sets according to this model of sample size N. Then we generate MAR missing data in each data set. Using the multiple imputation utilities in Mplus, see Asparouhov and Muthén (2010), we create 5 imputed data sets for each simulated data set. The imputed data sets are then analyzed with the true model and the two test statistics T_{imp} and $\bar{T}$ are computed. Since we analyze the data with the true model we expect the test statistics to accept, i.e., not reject the model, or more specifically to reject the model at the nominal level of 5%.

The data is generated using the following parameter values: $\mu_j = 0$, $\lambda_j = 1$, the factor variances and the residual variances are set to 1 and the correlation between the two factors is set to 0.5. We generate data sets of 3 different sample sizes $N = 100, 500$ and 1000. The missing data is generated according to the following missing data mechanism. For $j > 3$ the variable Y_j has no missing values, while for $j = 1, ..., 3$ the missing values are generated using the following formula

$$P(Y_j \text{ is missing}) = Exp(\alpha + \beta Y_{j+3})/(1 + Exp(\alpha + \beta Y_{j+3})), \qquad (4)$$

i.e., in this missing data mechanism the missing values for Y_j are predicted by another indicator variable Y_{j+3}. We use two different sets of parameters α and β to generate two different levels of missing data. Using $\alpha = -1.5$ and $\beta = 1$ we get approximately $L = 25\%$ of missing data for the variables Y_j, $j = 1, ..., 3$. Using $\alpha = -1$ and $\beta = 2$ we get approximately $L = 40\%$ of missing data for these variables.

Table 1 contains the average test statistic values and the rejection rates

Table 1: Average value (rejection rate) of different LRT test statistics in factor analysis model estimated with imputed data.

N	L	$\bar{T}$	T_{imp}
100	25%	18.0(.45)	9.2(.12)
500	25%	16.2(.45)	7.8(.08)
1000	25%	15.7(.46)	8.1(.05)
100	40%	26.5(.90)	18.8(.15)
500	40%	25.9(.86)	8.7(.09)
1000	40%	25.5(.78)	8.3(.09)

for both test statistics T_{imp} and $\bar{T}$. The degrees of freedom for this test is 8 so we expect to see average test statistic value of 8 if the test statistic works correctly. The results clearly indicate that the naive statistic $\bar{T}$ does not work correctly. It overestimates the test statistic value and underestimates the P-value. As a consequence this naive test statistic leads to inflated rejection rates. The more missing data the worse the performance of that test statistic. On the other hand the statistic T_{imp} appear to perform correctly in all cases. The average test statistic value is close to 8 and the rejection rate near the nominal level. It is clear also from the results that T_{imp} performs worst when the sample size is small $N = 100$ and there is a large portion of missing data, $L = 40\%$ for the first 3 variables. This is expected since T_{imp} is an asymptotic statistic. In that case the rejection rate is slightly inflated. The average test statistic value is also inflated, however that is mostly due to a single odd replication. In this simulation study we also encountered computational problems for T_{imp}. When $N = 100$ and $L = 25\%$ we had 1 replication with T_{imp} computational problems out of 100 replications. When $N = 100$ and $L = 40\%$ we had 9 replications with T_{imp} computational problems out of 100 replications. Thus we can conclude that small sample sizes and large amount of missing data are also the causes for T_{imp} computational problems.

Overall we conclude that T_{imp} yields an effective way to conduct LRT testing when we use multiple imputation data.

4 A power study

In this section we evaluate the power of the T_{imp} statistic to reject an incorrect model and we compare that to the power of the usual chi-square test statistic obtained by the FIML (full information maximum likelihood) estimator which we denote by T_{FIML}. The simulation study is conducted as in the previous section with the only change that the correlation between the two factors is set to 0.8. To evaluate the power of the chi-square statistics we use an incorrect model to analyze the data, we use a one factor analysis model. Thus we expect the chi-square statistics to reject the incorrect model particularly when the sample size is large. Tables 2 and 3 contain the rejection rates for the two test statistics for sample sizes $N = 100, 150, 200, 250$ and 300. Table 2 contains the results when the missing data is $L = 25\%$ while table 3 contains the results when the missing data is $L = 40\%$.

It is clear from these results that T_{FIML} is slightly more powerful than T_{imp}. The loss of power in T_{imp} appears to be increasing as the amount of missing data increases. Overall this loss of power appears to be small.

Table 2: Power study results for 25% missing data case. Percentage rejection rate.

N	300	250	200	150	100
T_{imp}	85	75	68	53	34
T_{FIML}	92	86	76	60	50

Table 3: Power study results for 40% missing data case. Percentage rejection rate.

N	300	250	200	150	100
T_{imp}	69	51	44	32	30
T_{FIML}	84	75	55	52	40

In addition T_{imp} has the advantage that the imputed missing data may be imputed from data that is not used in the estimated model and therefore can carry more information than the unimputed data set. This can lead to more accurate estimation and testing.

5 References

Asparouhov T. & Muthén B. (2010). Multiple Imputation with Mplus. Technical Report. www.statmodel.com

Enders, C.K. (2010). Applied missing data analysis. New York: Guilford.

Meng, X.L. & Rubin, D.B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. Biometrika, 79, 103-111.