Practical Approaches to Dealing with Nonnormal and Categorical Variables

Definitions and Distinctions

First, it is important to distinguish between categorical variables and continuous variables. Categorical variables are those with two values (i.e., binary, dichotomous) or those with a few ordered categories. Examples might include gender, dead vs. alive, audited vs. not audited, or variables with few response options like "never," "sometimes," or "always." Continuous variables are variables measured on a ratio or interval scale, such as temperature, height, or income in dollars.

Ordinal variables with many categories, such as 7-point Likert-type scales of agreement, are usually safely treated as "continuous." In practice, most researches treat ordinal variables with 5 or more categories as continuous, and there is some evidence to suggest this is not likely to result in much practical impact on results (e.g., Babakus, Ferguson, & Jöreskog, 1987; Dolan, 1994; Johnson & Creech, 1983; Hutchinson & Olmos, 1998). If ordinal variables with many categories are nonnormal, then data analytic techniques for nonnormal continuous variables should be used (see below). See Finney and DiStefano (2013) for a good summary of the justification of this general strategies in structural equation modeling.

When variables are measured on an ordinal scale and there are relatively few categories, 2-4 categories, estimation methods specifically designed for categorical variables are recommended. This includes nominal binary variables, because binary variables can be considered ordinal for the purpose of meaningful comparisons between the two groups (e.g., gender). A categorical analysis approach will have the greatest advantage (less bias) compared with standard ML when the following conditions hold: (1) when the values between categories are not equidistant; (2) when the relationship between the categorical measured variable and the theoretical variable it is supposed to measure is not a linear relationship—another way of stating (1); (3) when the ordinal variable is skewed or kurtotic.

Detection of Multivariate Nonnormality

So, how do you know your data are multivariate normal? The first step is to carefully examine univariate distributions and skew and kurtosis. West, Finch, & Curran (1995) recommend concern if skewness > 2 and kurtosis > 7. Kurtosis is usually a greater concern than skewness. If the univariate distributions are nonnormal, then the multivariate distribution will be nonnormal. One can have multivariate nonnormality (i.e., the joint distributions of all the variables is a nonnormal joint distribution) even when all the individual variables are normally distributed (although this is probably relatively infrequent in practice, at least not severe multivariate nonnormality). Therefore, one should also examine multivariate kurtosis and skewness. However, tests of multivariate normality are only available in EQS and LISREL. Mardia's multivariate skewness and kurtosis tests are distributed normally (z-test) in very large samples, so can be evaluated against a t, z, or chi-square distribution. Lawrence DeCarlo (1997) has developed macros for SPSS and SAS to calculate Mardia's multivariate skewness and kurtosis estimates and test them for significance (available at http://www.columbia.edu/~ld208/). EQS also provides a "normalized estimate" of Mardia's kappa. Bentler and Wu (2002) suggest that a normalized estimate greater than 3 will lead to chi-square and standard error biases.

Recommendations for Continuous Nonnormal Variables

In practice, some structural equation models with continuous variables (and generally including ordinal variables of five categories or more) will not have severe problems with nonnormality. The effect of violating the assumption of nonnormality is that chi-square is too large (so too many models are rejected) and standard errors are too small (so significance tests of path coefficients will result in Type I error).

The scaled chi-square and "robust" standard errors using ML estimation is a method suggested by Satorra and Bentler (1988; 1994). It appears to be a good general approach to dealing with nonnormality (Hu, Bentler, & Kano, 1992; Curran, West, & Finch, 1996). Adjustments are made to the chi-square (and to relative fit indices in some packages, such as Mplus, lavaan, and EQS) and standard errors based on a weight matrix derived from an estimate of multivariate kurtosis. Mplus prints this kurtosis adjustment,

referred to as the “scaling correction factor” (scf; or as d in Finney & DiStefano, 2013). The scaling correction factor is the standard chi-square divided by the scaled chi-square. The ratio is derived from a multivariate kurtosis estimate used to adjust the chi-square and standard errors. When data are multivariate normal, this scaling correction factor is 1.0, and there is no adjustment to the standard ML chi-square. The more multivariate kurtosis, the larger this scaling correction factor will be (e.g., 1.6 suggests the ML chi-square is approximately 60% higher than the scaled chi-square). At this point, no one has suggested a conventional value for the scaling correction factor that would indicate problematic levels of nonnormality, but I would be more concerned when the chi-square inflation is greater than 5 or 10% (scf of 1.05 or 1.10).

Depending on the complexity of the model and the severity of the problem, sample sizes of greater than 250 may be needed (Hu & Bentler, 1999; Yu & Muthén, 2002). For smaller samples, there is a potential danger of overcorrection with this method. This approach is available in LISREL (ML Robust), EQS (ML Robust), Mplus and lavaan (MLM for "maximum likelihood mean adjusted").

Bootstrapping is an increasingly popular and promising approach to correcting standard errors, but it seems that more work is needed to understand how well it performs under various conditions (e.g., specific bootstrap approach, sample sizes needed). The simulation work that has been done (Fouladi, 1998; Hancock & Nevitt,1999; Nevitt & Hancock, 2001) suggests that, in terms of bias, a standard “naïve” bootstrap seems to work at least as well as robust adjustments to standard errors. However, the Nevitt and Hancock (2001) results suggest that standard errors may be erratic for sample size of 200 or less and samples of 500 to 1,000 may be necessary to overcome this problem. The complexity of the model should be taken into account as their simulations were based on a moderately complex factor model (i.e., smaller sample sizes may be acceptable for simpler models). An alternative bootstrapping approach, the Bollen-Stine bootstrap approach, is usually recommended for estimation of chi-square. The Bollen-Stine chi-square approach seems to adequately control Type I error but there is some cost to power (Nevitt & Hancock, 2001). Bootstrapping approaches have now been incorporated in most major SEM packages.

Recommendations for Categorical Variables

There seems to be growing consensus that the best approach to analysis of categorical variables (with few categories) is the DWLS approach implemented in Mplus. This approach, usually referred to as a robust weighted least squares (WLS) approach in the literature (estimator = WLSMV or WLSM in Mplus and lavaan). The WLSMV approach seems to work well if sample size is 200 or better (Bandalos, 2014; Flora & Curran, 2004; Muthén, du Toit, & Spisic, 1997; Rhemtulla, Brosseau-Liard, & Savalei, 2012). In Lisrel and EQS, a similar approach that uses WLS together with polychoric correlations and asymptotic covariance matrices is used. In Mplus, there are two versions of DWLS estimates that have different approaches to setting the scaling of the y^* distribution, delta parameterization and theta parameterization. Delta parameterization (marginal) is the default and sets the scaling by setting the measurement residual to 1.0 and theta parameterzation (conditional) sets the scaling of the y^* variance to 1.0, estimating the measurement residual variance. Both can be called variants on the probit model, but theta parameterization corresponds more exactly to the probit regression estimates. These scaling choices are arbitrary in the sense that the chi-square for the model and the significance tests of the parameter estimates will be equal. DWLS works well in many situations (although one exception may be when missing data are MNAR), but a special full maximum likelihood estimation for binary or ordinal data can also be applied successfully. The full maximum likelihood estimation with categorical variables provides logit estimates. This approach is not yet available in many SEM software programs, but Mplus has implemented this when ESTIMATOR=ML is used in conjunction with dependent variables identified as categorical on the CATEGORICAL statement (a robust version of ML for binary data, which uses robust standard error estimates, is called MLR in Mplus). Probit and logistic estimates will often be quite similar in terms of their statistical conclusions. Work by Bandalos (2014) indicates that robust MLR performs better than the unadjusted ML and that MLR performed similarly to the WLSMV method. Compared with WLSMV, MLR has somewhat less power but better control of Type I error in smaller samples. Bandalos's work also suggests that sample sizes of 150 may be too small with either method, especially where distributions of the categorical variables are asymmetric.

In the most recent editions of Amos, an alternative approach to categorical variables has been added (Lee, 2007). The Bayesian approach requires an iterative process known as the Markov Chain Monte Carlo (MCMC). To date, there is fairly scant information on the performance of this approach with SEM with respect to fit estimation, the optimal algorithms to use, and standard errors under various conditions (cf. Lee & Yang, 2006). The Bayesian estimation process involves some artful judgment in the testing process. The Bayesian structural modeling approach has not become a popular alternative thus far (see Kaplan & Depaoli, 2012 for an introduction).

Fit Indices and Nested Tests

Relatively little simulation work on alternative fit indices (e.g., RMSEA, IFI, CFI) derived from robust approaches to nonnormal continuous variables (Satorra-Bentler robust approach or bootstrapping) is currently available, but the Satorra-Bentler scale chi-square appears to outperform the maximum unadjusted likelihood chi-square when data are nonnormal (Curran et al., 1996). Thus far, studies suggest that at least some alternative fit indices (TLI, CFI, RMSEA) using standard cutoffs (Hu & Bentler, 1999) also perform fairly well with the robust approach as long as the Satorra-Bentler scale chi-square is used to compute incremental fit indices and sample size is reasonably large (N = 250 or larger; Nevitt & Hancock, 2000; Yu & Muthén, 2002). The user should use some caution, because programs do not always recalculate incremental fit indices such as the CFI, TLI, or the IFI using the scaled chi-square for the tested model or the null model (I know that Mplus and EQS do use the scaled chi-squares in their calculation). Relative fit indices will likely be problematic when scaling corrections to the null model are not used (Hu & Bentler, 1999).

Perhaps less in known about how fit indices perform with DWLS under various circumstances—certainly not with the same level of precision on which Hu and Bentler based their recommendations about fit with continuous variables. The robust WLSMV chi-square used by Mplus seems to perform pretty well (Flora & Curran, 2004), although there is still likely to be a practical problem with using chi-square as a sole measure of fit because of its sensitivity to sample size. There is some evidence that RMSEA, TLI, and CFI perform reasonably well with categorical model estimation (DWLS; Beauducel & Herzberg, 2006; Hutchinson & Olmos, 1998; Yu & Muthén, 2002), and they are likely to perform best when robust adjustments are made to the chi-square.

The Weighted Root Mean Square Residual (WRMR) is a measure that Muthén has recommended for fit of models with categorical observed variables. Yu and Muthén (2002) recommend that a model with a WRMR of less than 1.0 indicates good fit (I have also seen the value of .9 recommended). In my experience the WRMR does not always give sensible results, and I do not recommend it as a sole indicator of fit with DWLS estimation. The Bayesian Information Criterion (BIC) is sometimes suggested as a measure of fit for categorical models, but there is no consistently used cutoff for good fit and the BIC may be most practical for comparing fit of different models. In sum, my best recommendation (consistent with the Finney & DiStefano, 2013, review) for evaluating model fit with the DWLS approach at this point in time is to use the TLI or CFI ($\geq$ .95) and RMSEA ($\leq$ .05), possibly in conjunction with the WRMR (approximately less than 1).

Nested tests (likelihood ratio test) require special attention for robust estimation. The scaling correction factor (scf) must be used to weight the difference (Satorra, 2000; Satorra & Bentler, 2001). The following formula gives the adjustment to the diffence in chi-square which can be used for significance testing:

$$\Delta\chi^2_{SB} = \frac{\chi^2_{M0} - \chi^2_{M1}}{\left(df_{M0}scf_{M0} - df_{M1}scf_{M1}\right)/df_{M0} - df_{M0}}$$

The scf is equal to the ratio of traditional ML chi-square to the Satorra-Bentler scale chi-square for the model, or $\text{scf} = \chi^2_{ML} / \chi^2_{SB}$.

Nested tests for ordinal analysis methods are not widely available in software programs currently, and there has been limited simulation work comparing methods. One suggestion has been to use a WLS estimator just for comparison of model chi-square values using the simple difference for chi-square and degrees of freedom. An alternative is a more elaborate vanishing tetrad test (Hipp & Bollen, 2003). Asparouhov and Muthén (2006) have adapted the tests developed by Satorra (2000) and Satorra and Bentler (2001) that computes the estimated ratio of the weighted likelihoods of two models using WLSMV estimation for ordinal variables. Mplus provides automated nested tests with the DIFFTEST command that can be used for several estimation or robust methods.

Missing Data Estimation with Non-normal or Categorical Data

For nonnormal continuous data where some data are missing, a variation on the full maximum likelihood can be used (Yuan-Bentler, 2006) approach and missing data with categorical. In Mplus and lavaan, this is obtained with ESTIMATOR = MLR. Missing data with categorical variables is best handled with full maximum likelihood (logistic estimation), which is only available currently in Mplus. WLSMV estimates with missing data may work well generally (Asparouhov & Muthén, 2006), but it is not a full information method and may not perform as well if data are not at least MAR. Amos, which offers FIML for missing data and bootstrapping for nonnormal data, does not currently have a method of dealing with both missing and nonnormal data simultaneously.

References

Asparouhov, T., & Muthén, B. (2010). *Weighted least squares estimation with missing data*. Unplublished technical report, retrieved from http://www.statmodel2.com/download/BayesAdvantages18.pdf.

Babakus, E., Fergnson, C. E. & Joreskog, K G, (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research, 24*, 2228.

Bentler, P.M., & Wu, E.J.C. (2002). *EQS for Windows user's guide*. Encino, CA. Multivariate Software, Inc.

Curran, P. J., West, S. G, & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*, 16-29.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*, 292-307.

Dolan, C. V. (1994), Factor analysis of variables with 2, 3, 5, and 7 response categories: A comparison of categorical variable estimators using simulated data. *British journal of Mathematical and Statistical Psychology, 47*, 309-326.

Finney, S.J., & DiStefano, C. (2013). Non-normal and categorical data in structural equation modeling. In G.R. Hancock & R.O. Mueller (Eds.), *Structural equation modeling: A second course, 2nd Edition* (pp. 439–492). Charlotte, NC: Information Age Publishing.

Fouladi, R.T. (1998, April). *Covariance structure analysis techniques under conditions of multivariate normality and non-normality—modified and bootstrap based test statistics.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Hancock, G.R., & Nevitt, J. (1999). Bootstrapping and the identification of exogenous latent variables within structural equation models. *Structural Equation Modeling*, *6*, 394-399.

Hutchinson, S, R., & Olmos, A (1998). Behavior of descriptive fit indexes in confirmatory factor analysis using ordered categorical data. *Structural Equation Modeling: A Multidisciplinary Journal*, *5,* 344-364.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6(1*), 1-55.

Hu, L., Bentler, P.M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112*, 351-362.

Johnson, D.R., & Creech, J.C. (1983) Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review, 48*, 398-407.

Kaplan, D., & Depaoli, S. (2012). In R.H. Hoyle (Ed.). *Handbook of structural equation modeling* (pp., 650-673). New York: Guilford Press.

Lee, S-Y., & Tang, N.-S. (2006). Bayesian analysis of structural equation models with mixed exponential family and ordered categorical data. *British Journal of Mathematical and Statistical Psychology, 59*, 151–172

Lee, S-Y. (2007). *Structural Equation Modeling: A Bayesian Approach.* New York: Wiley.

Muthén , B.O, du Toit, S., & Spisic,, D. (1997). *Robust infoerince using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.* Unpublished manuscript.

Nevitt, J., & Hancock, G.R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling, 8,* 353-377.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological methods*, *17*, 354-373.

Satorra, A., & Bentler, P.M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *1988 Proceedings of the Business and Economic Statistics Section of the American Statistical Association*, 308-313.

Satorra, A., & Bentler, P.M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye and C.C. Clogg (eds.), *Latent Variable Analysis: Applications to Developmental Research* (pp. 399-419). Newbury Park: Sage.

Yu, C.-Y., & Muthén, B. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* Doctoral dissertation. Retrieved from http:// www.statmodel.com/download/Yudissertation.pdf

Recommended reading: Finney, S.J., & DiStefano, C. (2013). Non-normal and categorical data in structural equation modeling. In G.R. Hancock & R.O. Mueller (Eds.), *Structural equation modeling: A second course, 2nd Edition* (pp. 439–492). Charlotte, NC: Information Age Publishing.