Alternative Estimation Methods

ML

Remember that the usual approach to estimating fit and coefficients in SEM with continuous variables is the maximum likelihood (ML) approach. ML uses derivatives to minimize the following fit function:

$$F_{\text{ML}} = \log\left|\boldsymbol{\Sigma}(\boldsymbol{\theta})\right| + tr\left(\mathbf{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\right) - \log\left|\mathbf{S}\right| - (p+q)$$

The fit function generally captures the discrepancy between the observed covariance matrix $\mathbf{S}$ and the covariance matrix implied by the model, $\boldsymbol{\Sigma}$. See the handout from this class "Maximum Likelihood" for more detail. The ML estimator assumes that the variables in the model are (conditionally) multivariate normal (i.e., the joint distribution of the variables is distributed normally).[1]

GLS

Generalized least squares is an alternative fitting function. The GLS fit function also minimizes the discrepancy between $\mathbf{S}$ and $\boldsymbol{\Sigma}$, but uses a weight matrix for the residuals, designated $\mathbf{W}$.

$$F_{\text{GLS}} = \left(\frac{1}{2}\right) tr\left(\left\{\left[\mathbf{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta})\,\mathbf{W}^{-1}\right]\right\}^2\right)$$

Notice that this is a much simpler function (e.g., no logs), and it is clear that the discrepancy between the obtained covariance matrix and the covariance matrix implied by the model $(\mathbf{S} - \boldsymbol{\Sigma})$ is minimized after weighting it by $\mathbf{W}$. Although any $\mathbf{W}$ can be chosen for the weight matrix, most commonly, the inverse of the covariance matrix, $\mathbf{S}$, is used in SEM packages. F_{GLS} is asymptotically equivalent to F_{ML}, meaning that as sample sizes increase, they are approximately equal. F_{GLS} is based on the same assumptions as F_{ML} and would be used under the same conditions. GLS is thought to perform less well in small samples, however, so F_{ML} is usually chosen instead of F_{GLS}. The simplicity of the function, however, means that other weight matrices could be used in an attempt to correct for violations of distributional assumptions.

ADF/WLS/AGLS

The asymptotic distribution free function for SEM is given by Browne (1984). It is described as arbitrary generalized least squares (AGLS) by Bentler in the EQS package and weighted least squares (WLS) by Jöreskog and Sörbom in LISREL (and the related approach described below used by Mplus and `lavaan`). The main advantage of the ADF estimator is that it does not require multivariate normality. The ADF estimator is based on the F_{GLS}, except a different $\mathbf{W}$ is chosen. It can be written in a general form that encompasses GLS, ML, and ULS (not discussed here) where the difference depends on the choice of $\mathbf{W}$:

$$F_{\text{ADF}} = F_{\text{AGLS}} = F_{\text{WLS}} = (\mathbf{s} - \boldsymbol{\sigma})'\,\mathbf{W}^{-1}(\mathbf{s} - \boldsymbol{\sigma})$$

$\mathbf{W}$ used in F_{ADF} is based on a covariance of all of the elements of the covariance matrix, $\mathbf{S}$. That is, a covariance matrix is constructed that estimates the covariances between each s_{ij} element of $\mathbf{S}$, and is therefore a ½[v(v+1)] by ½[v(v+1)] matrix, with v as the number of observed variables. It is important to realize that the "covariances of covariances" are related to kurtosis estimates (so called "fourth-order moments").[2] So, the GLS fit function is weighted by variances and kurtosis in an attempt to correct for violations of the normality assumption. Another way of saying this is that when the data are normal, the ADF estimator reduces to GLS because there is no kurtosis. The large weight matrix causes serious practical difficulties when there is a large number of variables in the model (e.g., more that 20 or so), and computer packages (e.g., EQS) do not allow estimation unless the number of cases is equal or greater than number of elements in the weight matrix (i.e., ½[v(v+1)] times ½[v(v+1)] divided by 2). Simulation studies suggest that chi-square values are severely overestimated with small samples and that sample sizes of about 5000

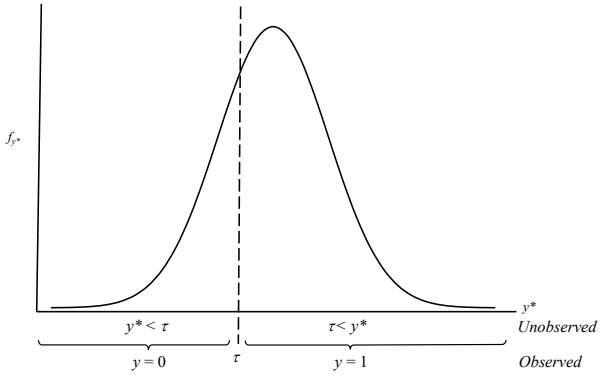
[1] The conditional portion of this assumption is that the distribution of the y variable that is of importance is the residual distribution. That is, if after accounting for the predictors of y, the distribution is normal, then the assumption is met. It also should be noted here that, like regression, there is not an assumption about the distribution of the predictor/independent variable, only the dependent variable.

[2] A raw form equation for univariate kurtosis is $\left[\sum(Y - \bar{Y})^4 / N\right] / s^4$, where the deviations from the mean are raised to the fourth power.

are necessary for good estimates. A study by Olsson, Foss, Troye, and Howell (2000) suggests that ADF estimation performs poorly when the model is misspecified. Combined with the practical limitation of the number of variables, this is usually seen as an unattractive approach for nonnormal or categorical variables.

WLS for Categorical Variables

The ADF estimator is not very practical as a general estimation approach in its original form, but it has been implemented with considerable success with categorical (binary and ordinal) variables in modified form. Models with categorical variables are always considered to be in violation of the normality assumption and, thus, the usual F_{ML} estimator is not recommended. In the context of categorical variable estimation, the ADF estimator is most often referred to these days as WLS. The modified approach that has developed is a multiple-step estimation involving polychoric correlations as input to create the asymptotic covariance matrix used for weighting in the WLS estimation. The idea behind the method is that categorical variables can be conceived of as having an underlying continuous unobserved variable, called y^*. y^* is estimated by *polychoric* correlations which correct for loss of information when Pearson correlations are used due to cruder categorization of a continuous variable (see Olsson, 1979; MacCallum, Zhang, Preacher, & Rucker, 2002). *Tetrachoric* correlations are a special case of polychoric correlations involving only binary variables, and *polyserial* correlations are those involving the correlation between a binary and a continuous variable. Often all three types are referred to more generally as polychoric correlations. The concept of y^* is the same as that invoked to conceptualize probit analysis. The variable y^* is a true value that is not observed but leads to the observed response of y, which is binary or ordinal. The value of y^* can be thought of as a propensity to respond 0 or 1 on the y variable in the case that y is binary, for example. The figure below is an analogue representation of the idea:



In a sense, the WLS estimation is performed on the estimated y^* variables in this two-step estimation process. LISREL requires the user to implement this process explicitly in separate steps, but other software programs, such as Mplus, lavaan, and EQS, allow the two-step process to be handled automatically (as long as raw data are available). The approach requires an inversion of the full weight matrix, which can become cumbersome when there are many variables. Estimation using this approach performs reasonably well statistically in a number of circumstances, but can be improved upon.

DWLS

Muthén (1978, 1993, 1984) suggested a modification of this general categorical variable approach, known as diagonal weighted least squares (DWLS; sometimes “diagonally”) estimation or a "limited information" approach. The DWLS approach uses the WLS estimator with polychoric correlations as input to create the asymptotic covariance matrix. The approach is computationally more practical because it avoids inversion of the large weight matrix (using “Taylor expansion” to obtain the asymptotic covariance matrix; Muthén, 1993; Muthén, du Toit, & Spisic, 1997). The method seems to perform better statistically as well (Distefano & Morgan, 2014; Rhemtulla, Brosseau-Laird, & Savalei, 2012), outperforming full WLS in small samples. The approach is typically paired with robust estimation adjustments (the same as the Satorra-Bentler robust estimates used with ML for nonnormal continuous variables) to improve standard errors, chi-square, and fit indices. In Mplus (as well as `lavaan`, and sometimes more generally in the literature), the DWLS with

adjustment is referred to as WLSMV.[3] The robust DWLS methods seem to work well in many conditions, including smaller samples and with nonnormal data (e.g., Bandalos, 2014; Rhumtulla et al., 2012). EQS (Bentler & Wu, 2002) uses an alternative robust method described as a "partitioned maximum likelihood" approach, obtaining estimates in separate steps depending on the types of variables involved.

The DWLS estimation produces probit regression estimates but there are two parameterizations—ways to scale the y^* distribution. Because it is not possible to estimate both the variance of measurement residual (θ_ε) and the variance of the y^* distribution at the same time, a scaling constraint must be made to either the measurement residual variance or $\text{var}(y^*)$. When $\text{var}(y^*)$ is assumed to be equal to 1 (as is in probit regression; Long, 1997), the parameterization method is called *theta*. For theta parameterization, the measurement residual is freely estimated and equality constraints or correlated residuals are possible. When the measurement residual is assumed to be 1 and $\text{var}(y^*)$ is estimated, the parameterization is called *delta*. With delta parameterization, values of $\text{var}(y^*)$ can be set or constrained to be equal to the $\text{var}(y^*)$ of other variables.

Marginal Maximum Likelihood
A maximum likelihood method (sometimes also maximum marginal likelihood) for dichotomous and ordinal variables, which has been commonly used in Item Response Theory (IRT) modeling (Bock & Aitkin, 1981), is another available estimation method available for SEM estimation in Mplus (and for confirmatory factor models in `lavaan`) when variables are declared as ordinal and ML estimation is requested.[4] The approach, which also has a history in factor analysis (Christofferssen, 1975; Muthén, 1978; Muthén & Christofferssen, 1981) produces logistic parameter estimates by default (although probit estimates can be requested) and Pearson chi-square and likelihood ratio chi-squares for fit (based on multiway frequency table). An adaptive quadrature integration method is used for approximation of the parameter distribution and is computationally intensive for larger models. In comparisons with DWLS, the marginal maximum likelihood method performs comparably well when paired with robust adjustments (Bandalos, 2014; Demars, 2012; Newsom & Smith, 2020). One potential advantage is that it is a full information maximum likelihood method that only assumes MAR, whereas DWLS has stricter missing data assumptions (MCAR; or MAR when only x variables can affect missingness; Asparouhov & Muthén, 2010b).

The Satorra-Bentler Scaled Chi-square and Standard Errors
Satorra and Bentler suggested that multivariate kurtosis estimates be used to “scale” or correct the chi-square value and standard errors for continuous nonnormal variables (Satorra & Bentler, 1988; Satorra & Bentler, 1994). The corresponding approach with missing data was described by Yuan and Bentler (2000). Chi-square is usually inflated with nonnormal samples and standard errors are usually too small (although depending on whether the distribution is platykurtic or leptokurtic, the direction of bias can differ). This approach is used for continuous nonnormal variables and appears to do fairly well with small samples (200-500 cases; Curran, West, & Finch, 1996). This is the same robust adjustment used for in conjunction with DWLS for categorical variables.

Bootstrapping
Bootstrap is another approach to problems with nonnormality (but is not typically recommended for binary and ordinal variables with few categories). In the bootstrap approach, a large number of samples (usually a minimum of 500 or 1000 are recommended) are drawn from your data. The samples are drawn *with replacement,* so that the same cases may be drawn into the same bootstrap sample. These repeated samples create a mini sampling distribution, and based on the central limit theorem, it should have desirable distributional characteristics. There are a number of variations on bootstrapping with SEM, including “naïve” bootstrap, bias correction, and bias corrected accelerated (but see Bollen & Stine, 1993; Yung & Bentler, 1996). The bootstrap samples are used to calculate new standard errors ("naïve" bootstrap) and can be used to correct the chi-square for fit (Bollen-Stine bootstrap). The z-tests or “critical ratio” uses the bootstrap standard errors, and are considered “approximate” significance tests. Evidence by Nevitt and Hancock (2001) suggest an original sample size of 500 or greater may be needed for stable bootstrap estimation.

[3] Another estimation option in Mplus, called WLSM, uses just the means (rather than means and variances) for adjustments, but this method is known to have inflated rejection rates (Asparouhov & Muthen, 2010a).
[4] See Demars (2010) and Bauer and Kamata (2008) for explanations of the equivalence of categorical confirmatory factor analysis and IRT.

Bayesian SEM

A Bayes analysis approach has become increasingly available in structural equation packages (e.g., in WinBUGS/OpenBUGS, Mplus, the R package `blavaan`, Amos) for analysis of continuous and categorical data, although it has yet to become frequently employed by researchers. Bayesian analysis can make less strict assumptions about sampling distributions, and, thus, it has potential advantages in circumstances in which traditional SEM estimation might not work well. The Bayesian estimation method is theoretically, fundamentally different from maximum likelihood or other classical significance testing ("frequentist") approaches. Bayesian estimation uses prior information about the distribution to seek a posterior probability of the parameter(s) given the data, $p(\theta|y)$.

$$p(\theta \mid y) = \frac{p(\theta)\, p(y \mid \theta)}{p(y)}$$

The equation follows Bayes' Theorem, where θ is a random parameter (or set of parameters) and y represents a random variable (or set of variables). In a way, the likelihood of the parameter is weighted by prior expectations. Maximum likelihood takes a different theoretical approach in that the parameter is estimated by fitting the observed values to specific expected values (either a null hypothesis or one implied by the model) seeking the best fit possible, or seeking to find the likelihood of a certain parameter value given the data, or $L(\theta|y)$. No prior information is taken into account. With Bayesian information, prior information can increase the probability of finding a true parameter value if it offers correct hints as to the true parameter values. If prior information is informative about the parameter of interest (e.g., normal distribution with a specific mean and variance) finding the correct parameter value will be more likely if the prior distributional information is correct. If the prior information about the distribution is uninformative (e.g., uniform distribution with infinite bounds), then there is no additional advantage over a classical approach. If prior information is incorrect or misleading, Bayesian analysis can lead to erroneous conclusions (e.g., Diaconis & Freedman, 1986).

Hypothesized models are specified in the usual manner, but location (expected value), variance, and distributional information can be submitted as priors. The computational approach then uses a Markov chain Monte Carlo (MCMC) iterative sampling process to identify the probable parameter value. The process derives an empirical range of possible values (in a not-dissimilar way to bootstrapping), resulting in posterior estimates of about the expected parameter value and its distribution. Because of the different theoretical orientation of Bayesian analysis, the estimate intervals are considered "credibility intervals" or "credible intervals" in which the interval is expected to contain the correct parameter value. A number of decisions are needed by the researcher in addition to specifying priors, including starting values (specific or randomly generated), the number of iterations, the number of MCMC chains (streams of iterative sampling), the number of burn-in samples (initial samples discarded), the number of iterations, and the checking processes used to see whether "convergence" has concluded appropriately. See Kaplan & Depaoli (2012) and Depaoli (2021) for good introductions to Bayesian SEM and Asparouhov and Muthén (2010c; 2021) for more computational and implementation details in Mplus.

Other Estimators

Other possible estimators include two-stage least squares (2SLS), three-stage least squares (3SLS), ordinary least squares (OLS), and unweighted least squares (ULS). Most of these approaches are seldom used, because they provide poor estimation (e.g., ULS) or because they have not been very thoroughly investigated (e.g., 3SLS). 2SLS has received more attention in statistical papers because it does not rely on normality assumptions and is one approach to moderator tests in SEM (more on this topic later; see Bollen & Biesanz, 2002; Bollen & Paxton, 1998).

Summary and Recommendations

These are the most commonly used estimators and generally supported by simulation studies as producing good fit statistics and standard errors generally for each of the types of data listed. See "SEM with Nonnormal Continuous Variables" for more information and more references.

Most Commonly Used and Generally Recommended Estimation Settings

Data	Mplus & lavaan Estimation Settings	Notes
MVN, no missing	ML	Maximum likelihood MVN data assumed no missing values declared for any dependent variables used in the model
MVN, missing	ML "ml" or "fiml" in lavaan	Full information maximum likelihood (FIML). Default whenever missing values are identified for any dependent variables included in the model
Nonnormal continuous variables, no missing	MLM	Maxmimum likelihood mean adjusted. Satorra-Bentler chi-square and robust standard errors
Nonormal continuous variables, missing	MLR	Yuan-Bentler adjusted chi-square and robust standard errors
Binary or ordinal variables	WLSMV	Also called diagonal weighted least squares (DWLS). WLSMV stands for weighted least squares mean and variance adjusted. Correct chi-square and robust standard method used in Satorra-Bentler approach. Default if any categorical ("ordinal" in lavaan) variables declared. Marginal maximum likelihood is an alternative with fewer assumptions about missing data mechanisms (MLR when any categorical variables are declared) but less useful information about fit.

MVN = multivariate normal.

References and Further Reading

Asparouhov, T., & Muthén, B. (2010a). *Simple second order chi-square correction.* Unpublished manuscript. https://www.statmodel.com/download/WLSMV_new_chi21.pdf
Asparouhov, T., & Muthén, B. (2010b). *Weighted least squares estimation with missing data.* Unplublished technical report, https://www.statmodel.com/download/GstrucMissingRevision.pdf
Asparouhov,T., & Muthén, B. (2010c). *Bayesian analysis in Mplus: Technical Implementation*. Unpublished technical report, https://www.statmodel.com/download/Bayes3.pdf.
Asparouhov, T., & Muthén, B. (2021, September). *Bayesian analysis of latent variable models using Mplus, version 5*. Retrieved from https://www.statmodel.com/download/BayesAdvantages18.pdf
Bandalos, D.L. (2014). Relative Performance of Categorical Diagonally Weighted Least Squares and Robust Maximum Likelihood Estimation. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(1), 102-116
Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443–459.
Bollen, K.A. (2001). Two-Stage Least Squares and Latent Variable Models: Simultaneous Estimation and Robustness to Misspecifications. Ch 7, pp 199-138 in R. Cudeck, S. du Toit and D. Sorbom (eds) Structural Equation Modeling: Present and Future: A Festschrift in honor of Karl Joreskog, Scientific Software International: Lincolnwood.
Bollen, K. A., & Biesanz, J. C. (2002). A note on a two-stage least squares estimator for higher-order factor analyses. *Sociological Methods & Research, 30*, 568-579.
Bollen, K. A., & Paxton, P. (1998). Two-stage least squares estimation of interaction effects. In R. E. S. G. A. Marcoulides (Eds.), Interaction and nonlinear effects in structural equation modeling (pp. 125-151). Mahwah, NJ, Lawrence Erlbaum Associates.
Bollen, K.A., & Stine, R.A. (1993). Bootstrapping goodness-of-fit measures in structural equation modeling. In K.A. Bollen & J.S. Long (Eds.), Testing structural equation models. Newbury Park: Sage.
Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40*, 5–32.
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.
Demars, C.E.. (2010, May). A Comparison of Limited-Information and Full-Information Methods in Mplus for Estimating IRT Parameters for Non-normal Populations. Paper to be presented at the annual meeting of the National Council on Measurement in Education, Denver.
DeMars, C. E. (2012). A comparison of limited-information and full-information methods in M plus for estimating item response theory parameters for nonnormal populations. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(4), 610-632.
Depaoli, S. (2021). *Bayesian structural equation modeling.* Guilford.
DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling: A multidisciplinary journal, 21*(3), 425-438.
Diaconis, P., & Freedman, D. (1986). On the consistency of Bayes estimates. *The Annals of Statistics, 14,* 1-26.
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.
Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal, 15*(1), 136-153.
Kaplan, D., & Depaoli, S. (2012). "Bayesian Structural Equation Modeling." In Hoyle, R. H. (Ed.), *Handbook of structural equation modeling* (pp. 651-673). New York: Guilford press.
Long, J.S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
MacCallum, R.C., Zhang, S., Preacher, K.J., & Rucker, D.D. (2002). On the practice of dichotimization of quantitative variables. Psychological Methods, 7, 19-40.
Muthén, B. O. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*, 551–560.
Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*, 115–132.
Muthén , B.O. (1993). Goodness of fit with categorical and other nonnormal variables. In K.A. Bollen, & J.S. Long (eds.), *Testing structural equation Models* (pp. 205-234). Newbury Park, CA: Sage.
Muthén, B., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 46*, 407–419.
Muthén , B.O, du Toit, S., & Spisic,, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.* Unpublished manuscript.
Newsom, J. T., & Smith, N. A. (2020). Performance of latent growth curve models with binary variables. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(6), 888-907.
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.
Olsson, u. H. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psycltometrika, 44*, 442-460.
Olsson, U.H., Foss, T., Troye, S. V., & Roy D. Howell (2000). The Performance of ML, GLS and WLS Estimation in Structural Equation Modeling Under Conditions of Misspecification and Nonnormality. Structural Equation Modeling, 7 (4), 557-595.
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.https://www.statmodel.com/download/Article_075.pdf.
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.
Yuan K. H. and P.M. Bentler. 2000. Three Liklihood-Based Methods for Mean and Covariance Structure Analysis with Non-Normal Missing Data. *Sociological Methodology 2000*,165-200. Washington, D.C. American Sociological.
Yung, Y.-F., & Bentler, P.M. (1996). Bootstrapping techniques in analysis of mean and covariance structures. In G. A. Marcoulides & R. E. Schumacker (Eds.), Advanced Structural Equation Modeling. Issues and Techniques (pp.125-157). Mahwah, NJ: Lawrence Erlbaum.