The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation

Model–Data Fit

Contributors: Brian C. Wesolowski
Edited by: Bruce B. Frey
Book Title: The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation
Chapter Title: "Model–Data Fit"
Pub. Date: 2018
Access Date: February 26, 2018
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks,
Print ISBN: 9781506326153
Online ISBN: 9781506326139
DOI: http://dx.doi.org/10.4135/9781506326139.n439
Print pages: 1079-1081

Empirical models play an important role in bringing order, comprehension, and manageability to complex interrelationships among variables. They enhance researchers' abilities to develop hypotheses and provide mechanisms to speculate about multifaceted processes. In educational and related psychological research, empirical models are most often developed in order to explain latent constructs and are therefore considered to be only approximations of reality. Latent constructs are inferred based upon observable (i.e., measured) indicators or behaviors, each subject to errors in measurement. Goodness-of-fit measures, used in the context of latent construct modeling, describe how well the observed data represents the latent constructs of interest. Inferences related to latent constructs are drawn from these observable occurrences; therefore, assessing the goodness-of-fit for a model is one of the most important aspects to the validity of interpretation in model building processes.

Beginning in the 1980s, methodologies for fit evaluation were rapidly and exhaustively conducted in educational and psychometric research, resulting in a multitude of approaches. First, this entry describes applications of basic goodness-of-fit tests. Second, this entry broadly surveys some of the more commonly used examples of absolute model fit indices that answer the question, "*Does the hypothesized model provide an overall fit to the observed data?*" Third, this entry broadly surveys some of the more commonly used examples of comparative model fit indices that answer the question, "*Which model most adequately replicates under different sample selections?*" Lastly, this entry provides a brief overview of parsimonious fit indices.

Goodness-of-Fit Statistics

Historically, early applications of fit evaluation included goodness-of-fit tests for observed categorical frequencies placed within a contingency table, where adequacy of fit was based upon the error variance of the model. However, when applied to linear measures or, more specifically, linear model building (i.e., factor analytic models including path analysis or structural equation modeling), error variance for collected observations is unknown. Therefore, traditional methods for fit evaluation are rendered not suitable.

In the context of linear model building, two of the more commonly used methods for evaluating overall model fit is the likelihood ratio chi-square goodness-of-fit statistic and the Pearson chi-square goodness-of-fit statistic. Use of these goodness-of-fit statistics came at the advent of the maximum likelihood estimation for the multinomial distribution. Using the maximum likelihood estimation procedure, the sampling distributions are based upon asymptotic distributions and use a vector of frequencies from nongrouped, observed data. It is important to note that a goodness-of-fit *statistic* (as compared to a goodness-of-fit *index* [GFI]) is a type of GFI with a known sampling distribution. Use of a goodness-of-fit *statistic* allows the researcher to conduct hypothesis testing for overall model fit. In particular, the chi-square goodness-of-fit statistic tests the hypothesis that the obtained population covariance input matrix of the observed data matches the model-implied covariance input matrix expected by the hypothesized model.

Traditionally, larger chi-square statistics, in relation to their degrees of freedom, indicate a lack of model-data fit. Smaller chi-square statistics, in relation to their degrees of freedom, indicate good model-data fit. In the application chi-square statistics to linear model building processes, researchers are not interested in rejecting the null hypothesis; rather, they are interested in accepting the null hypothesis where insignificant differences are desirable. In these instances, smaller chi-square values indicate good model-data fit. Therefore, a significant chi-square

statistic suggests that the model does not fit the data. Conversely, an insignificant chi-square statistic indicates adequate model-data fit. Furthermore, *p* values attached to the chi-squares with adequate model-data fit would be expected to demonstrate nonsignificance.

However, several weaknesses have been documented regarding the traditional chi-square statistic for use as a qualifier of true model adequacy. These weaknesses include violations to the assumption of multivariate normality and sensitivity to sample size and strength of correlations. This oversensitivity to model discrimination can often result in considerable Type I errors. Consequentially, the researcher may choose to move to an ad hoc measure of fit where transformations to the asymptotic chi-square statistic can provide more robust management of the observed data. These alternative measures can include but are not limited to the scaled chi-square statistic, the adjusted chi-square statistic, or the WLSMV chi-square estimator, for example. Furthermore, the chi-square is a measure of *exact* fit, which contradicts the conceptual notion that model building processes are based upon *approximations* of reality. Therefore, retaining the null hypothesis is never to be expected. As a result, the acceptance of the null hypothesis is not generally of interest to the researcher. Properties of goodness-of-fit *indices* are therefore more relevant and meaningful in the context of linear model building.

Goodness-of-Fit Indices

Goodness-of-fit indices can be classified into three broad categories of practical fit indices: (1) absolute fit indices, (2) comparative fit indices, and (3) parsimonious fit indices.

Absolute Fit Indices

Absolute fit indices determine the degree to which the hypothesized model predicts, or fits, to the observed data. These indices do not use an a priori baseline model for comparison. Rather, they provide a measure derived from the model fit of the observed and hypothesized covariance matrices. Absolute fit indices answer the question, "*Does the hypothesized model provide an overall fit to the observed data?*" Absolute fit indices assess the overall model fit of the hypothesized model using statistical hypothesis tests represented by one single statistical index. Absolute fit statistics answer the question, "*Overall, how well could the hypothesized model reproduce the observed data?*" The measure itself evaluates the magnitude of the discrepancy between the sample and model-estimated covariance input matrices.

In evaluating overall model fit, rejection of a null hypothesis is not necessarily informative. What is more interesting to the researcher is the magnitude and location of the misfit. One method of evaluating misfit is through an analysis of residuals. One example of an absolute fit measure that provides an index of residuals is the GFI and the closely related adjusted GFI. The GFI calculates the proportion of variance accounted for in comparing how much better the hypothesized model fits compared to no model. The GFI is calculated using the sum of squared residuals and sum of squared variances. The adjusted GFI is an adjustment to the GFI that uses the model's degrees of freedom. Conceptually, the relationship between the GFI and the adjusted GFI is similar to the relationship between R^2 and adjusted R^2 in the context of an ordinary least squares regression, where the model is adjusted based upon the amount of predictors in the model.

Another example of an absolute fit index that provides an index of residuals is the root mean square residual (RMR). The RMR is calculated as the square root of the difference between

the residuals of the obtained population covariance input matrix and the residuals of the model-implied covariance input matrix. However, the RMR is problematic, as the maximum value is unbound, resulting in difficulty of interpreting the acceptability of model-data fit. The RMR is also problematic, as the reported calculations are based upon the specific scale categories. As an example, if a measure does not provide similar categories for every item (i.e., a partial credit-type measure), the interpretation of results are unclear. Therefore, in these instances, the standardized RMR provides a more meaningful and substantive interpretation. However, the RMR and standardized RMR still do not provide specific information on where the misfit occurs, only a single index of residuals. Furthermore, both indices confound the error of sampling with the error of approximation.

The root mean square error of approximation, unlike the RMR and standardized RMR, simultaneously takes into account two potential sources of misfit: (1) error of approximation and (2) error of sampling. In doing so, the index is more robust to the centrality of the chi-square distribution and is independent of sample size. The error of approximation refers to the lack of fit of the hypothesized model to the population covariance matrix. The error of estimation refers to the closeness between the model-data fit of the sample and the model-data fit of the population. The parsing of both sources provides an index that simultaneously offers a measure of discrepancy between the obtained population covariance input matrix of the observed data and the model-implied covariance input matrix expected by the hypothesized model. The root mean square error of approximation index answers the question, "*How much is the error of approximation discrepant from the error of estimation due to sampling error?"*

Comparative Fit Indices

Comparative fit indices, also referred to in the research literature as incremental or relative fit indices, are a category of fit indices that compare the hypothesized model to some type of restricted, nested baseline (i.e., null) model. The null hypothesis and expectation of models for these indices are that the observed variables are uncorrelated, thereby not inferring evidence of a latent variable. In most cases, covariances between all input indicators are fixed to 0 in the baseline model. As a result of the overly severe constrainment, the baseline model is expected to demonstrate poor fit with large chi-square statistics. Comparative fit indices answer the question, "*Which model most adequately replicates under different sample selections?"*

One example of a comparative fit index (CFI) is the normed fit index (NFI) or the Bentler-Bonett NFI. The NFI compares the chi-square value (or fit function value) of the hypothesized model to the chi-square value (or fit function value) of the null model. A drawback of the NFI is its sensitivity to sample size. Small sample sizes often underestimate fit of the hypothesized model.

The Tucker Lewis index (TLI) overcame some of the limitations of the NFI. The TLI compares the mean square of the hypothesized model to the mean square of the null model. In some research literature, the TLI is referred to as the nonnormed fit index when discussed in the context of covariance structure analysis. The index can be represented as a proportion between the discrepancy between the hypothesized and null model. Limitations of the TLI/nonnormed fit index include a negative bias to smaller sample sizes, sensitivity to models more complex (i.e., more parameter estimates) in nature, and difficulty in interpreting the indices due to their nonnormed nature.

The incremental fit index (IFI), also referred to as DELTA2 in the research literature, was proposed as an improvement to the NFI. Specifically, the IFI adjusts for the NFI's sensitivity to small sample sizes by accounting for the hypothesized model's degrees of freedom. However, some drawbacks of the IFI include a positive bias to small sample sizes and a penalty for parsimony in the model due to the inclusion of the degrees of freedom for the hypothesized model.

One of the most reported comparative fit indices in research literature is the comparative fit index (CFI). The CFI was developed as an improvement to the NFI and IFI in that it is robust to small sample sizes. Conceptually similar to the logic of the root mean square error of approximation, the CFI measures improvements in noncentrality by fixing the noncentrality parameter to 0. As a result, estimation procedures are not affected by the sample size.

Parsimonious Fit Indices

In the context of linear model building, parsimony refers to the least amount of estimated parameters needed to achieve an adequate level of model-data fit. Conceptually, adding parameters to the model will improve model fit; however, adding the additional parameters may not be justified or warranted from a model fit perspective. Parsimonious fit indices provide a measure of discrepancy between the sample and model-estimated covariance input matrices while taking into consideration the complexity (i.e., the number of estimated parameters) of the model. Model parsimony favors more simple (i.e., less estimated parameters) hypothesized models over more complex (i.e., more estimated parameters) hypothesized models. Parsimony-corrected fit indices compare overidentified models with restricted modes and make adjustments to many of the indices previously described as a way to penalize for complexity of the model. Fit indices become lower the more complex the hypothesized model is, and generally, parsimonious fit indices have lower values of adequate model fit than other fit indices. Parsimony-corrected fit indices are proportion-based indices that are broadly calculated as the ratio of the number of degrees of freedom used by the model and the total number of degrees of freedom. Parsimonious fit indices adjust for losses in degrees of freedom by comparing an overfit model (i.e., excessive coefficients) with a restricted model. Examples of parsimonious fit indices include the parsimony GFU, parsimony NFI, Type 2 parsimonious NFI, the parsimonious CFI, and the Akaike information criterion.

See alsoChi-Square Test; Goodness-of-Fit Tests

Brian C. Wesolowski
http://dx.doi.org/10.4135/9781506326139.n439
10.4135/9781506326139.n439

Further Readings

Balakrishnan, N., Voinov, V., & Nikulin, M. S. (2013). Chi-squared goodness of fit tests with applications. Oxford, UK: Elsevier.
Bentler, P. M. (1990). Comparative fit indexes in structural models. Psychological Bulletin, 107(2), 238–246.
Bentler, P. M., & Bonnet, D. C. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88(3), 588–606.
Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. Structural Equation Modeling, 6(1), 56–83.
Hu, L. T., & Bentler, P. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), Structural equation modeling. Concepts, issues, and applications (pp. 76–99). London, UK: SAGE.

Mulaik, S. A., James, L. R., Alstine, J. V., Bennet, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. Psychological Bulletin, 105(3), 430–445.
Sivo, S. A, Fan, X., Witta, E. L., & Willse, J. T. (2006). The search for "optimal" cutoff properties: Fit index criteria in structural equation modeling. The Journal of Experimental Education, 74, 267–288.
Yuan, K. H. (2005). Fit indices versus test statistics. Multivariate Behavioral Research, 40(1), 115–148.