Goodness-of-Fit Testing

A Maydeu-Olivares and C García-Forero, University of Barcelona, Barcelona, Spain

Glossary

Absolute goodness of fit – The discrepancy between a statistical model and the data at hand.
Goodness-of-fit index – A numerical summary of the discrepancy between the observed values and the values expected under a statistical model.
Goodness-of-fit statistic – A goodness-of-fit index with known sampling distribution that may be used in statistical-hypothesis testing.
Relative goodness of fit – The discrepancy between two statistical models.

Introduction

The goodness of Fit (GOF) of a statistical model describes how well it fits into a set of observations. GOF indices summarize the discrepancy between the observed values and the values expected under a statistical model. GOF statistics are GOF indices with known sampling distributions, usually obtained using asymptotic methods, that are used in statistical hypothesis testing. As large sample approximations may behave poorly in small samples, a great deal of research using simulation studies has been devoted to investigate under which conditions the asymptotic p-values of GOF statistics are accurate (i.e., how large the sample size must be for models of different sizes).

Assessing absolute model fit (i.e., the discrepancy between a model and the data) is critical in applications, as inferences drawn on poorly fitting models may be badly misleading. Applied researchers must examine not only the overall fit of their models, but they should also perform a piecewise assessment. It may well be that a model fits well overall but that it fits poorly some parts of the data, suggesting the use of an alternative model. The piecewise GOF assessment may also reveal the source of misfit in poorly fitting models.

When more than one substantive model is under consideration, researchers are also interested in a relative model fit (i.e., the discrepancy between two models; see Yuan and Bentler, 2004; Maydeu-Olivares and Cai, 2006). Thus, we can classify GOF assessment using two useful dichotomies: GOF indices versus GOF statistics, and absolute fit versus relative fit. In turn, GOF indices and statistics can be classified as overall or piecewise. A third useful dichotomy to classify GOF assessment is based on the nature of the observed data, discrete versus continuous. Historically, GOF assessment for multivariate discrete data and that for multivariate continuous data have been presented as being completely different. However, new developments in limited information GOF assessment for discrete data reveal that there are strong similarities between the two, and here, we shall highlight the similarities in GOF assessment for discrete and continuous data.

GOF testing with Discrete Observed Data

Consider modeling N observations on n discrete random variables, each with K categories, such as the responses to n test items. The observed responses can then be gathered in an n-dimensional contingency table with $C = K^n$ cells. Within this setting, assessing the GOF of a model involves assessing the discrepancy between the observed proportions and the probabilities expected under the model across all cells $c = 1, \ldots, C$ of the contingency table. More formally, let π_c be the probability of one such cell and let p_c be the observed proportion. Let $\boldsymbol{\pi}(\boldsymbol{\theta})$ be the C-dimensional vector of model probabilities expressed as a function of, say, q model parameters to be estimated from the data. Then, the null hypothesis to be tested is $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$, that is, the model holds, against $H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}(\boldsymbol{\theta})$.

GOF Statistics for Assessing Overall Fit

The two standard GOF statistics for discrete data are Pearson's statistic

$$X^2 = N \sum_{c=1}^{C} (p_c - \hat{\pi}_c)^2 / \hat{\pi}_c, \qquad [1]$$

and the likelihood ratio statistic

$$G^2 = 2N \sum_{c=1}^{C} p_c \ln(p_c / \hat{\pi}_c). \qquad [2]$$

where $\hat{\pi}_c = \pi_c(\hat{\boldsymbol{\theta}})$ denotes the probability of cell c under the model.

Asymptotic p-values for both statistics can be obtained using a chi-square distribution with $C - q - 1$ degrees of freedom when maximum likelihood estimation is used. However, these asymptotic p-values are only correct when

all expected frequencies $N\hat{\pi}_c$ are large (>5 is the usual rule of thumb). A practical way to evaluate whether the asymptotic p-values for X^2 and G^2 are valid is to compare them. If the p-values are similar, then both are likely to be correct. If they are very different, it is most likely that both p-values are incorrect.

Unfortunately, as the number of cells in the table increases, the expected frequencies become small (as the sum of all C probabilities must be equal to 1). As a result, in multivariate discrete data analysis, most often, the p-values for these statistics cannot be trusted. In fact, when the number of categories is large (say k> 4), the asymptotic p-values almost invariably become inaccurate as soon as $n > 5$. To overcome the problem of the inaccuracy of the asymptotic p-values for these statistics, two general methods have been proposed: resampling methods (e.g., bootstrap), and pooling cells. Unfortunately, existing evidence suggest that resampling methods do not yield accurate p-values for the X^2 and G^2 statistics (Tollenaar and Mooijart, 2003). Pooling cells may be a viable alternative to obtain accurate p-values in some instances. For instance, rating items with five categories can be pooled into three categories to reduce sparseness. However, if the number of variables is large, the resulting table may still yield some small expected frequencies. Moreover, pooling may distort the purpose of the analysis. Finally, pooling must be performed before the analysis is made to obtain a statistic with the appropriate asymptotic reference distribution.

Due to the difficulties posed by small expected probabilities on obtaining accurate p-values for GOF statistics assessing absolute models, some researchers have resorted to examining only the relative fit of the models under consideration, without assessing the absolute model fit. Other researchers simply use GOF indices.

GOF Indices

With L denoting the loglikelihood, two popular GOF indices are Akaike's information criterion (AIC), AIC $= -2L + 2q$ and Schwarz Bayesian information criterion (BIC), BIC $= -2L + q\ln(N)$,

$$\text{AIC} = -2L + 2q, \text{BIC} = -2L + q\ln(N) \qquad [3]$$

The AIC and BIC are not used to test the model in the sense of hypothesis testing, but for model selection. Given a data set, a researcher chooses either the AIC or BIC, and computes it for all models under consideration. Then, the model with the lowest index is selected. Notice that both the AIC and BIC combine absolute fit with model parsimony. That is, they penalize by adding parameters to the model, but they do so differently. Of the two, the BIC penalizes by adding parameters to the model more strongly than the AIC.

GOF Statistics for Piecewise Assessment of Fit

In closing this section, the standard method for assessing the source of misfit is the use of z-scores for cell residuals

$$\frac{p_c - \hat{\pi}_c}{\text{SE}(p_c - \hat{\pi}_c)}, \qquad [4]$$

where SE denotes standard error. In large samples, their distribution can be approximated using a standard normal distribution. Unfortunately, the use of these residuals even in moderately large contingency tables, is challenging. It is difficult to find trends in inspecting these residuals, and the number of residuals to be inspected is easily too large. Most importantly, for large C, because the cell frequencies are integers and the expected frequencies must be very small, the resulting residuals will be either very small or very large.

New Developments in GOF with Discrete Observed Data: Limited Information Methods

In standard GOF methods for discrete data, contingency tables are characterized using cell probabilities. However, they can be equivalently characterized using marginal probabilities. To see this, consider the following 2×3 contingency table:

	$X_2 = 0$	$X_2 = 1$	$X_2 = 2$
$X_1 = 0$	π_{00}	π_{01}	π_{02}
$X_1 = 1$	π_{11}	π_{11}	π_{12}

This table can be characterized using the cell probabilities $\boldsymbol{\pi} = (\pi_{00}, \cdots, \pi_{12})'$. Alternatively, it can be characterized using the univariate $\dot{\boldsymbol{\pi}}_1 = (\pi_1^{(1)}, \pi_2^{(1)}, \pi_2^{(2)})$ and bivariate $\dot{\boldsymbol{\pi}}_2 = (\pi_{1\,2}^{(1)(1)}, \pi_{1\,2}^{(1)(2)})$ probabilities, where $\pi_i^{(k)} = \Pr(X_i = k)$, $\pi_{i\ j}^{(k)(l)} = \Pr(X_i = k,\ X_j = l)$, and

	$X_2 = 0$	$X_2 = 1$	$X_2 = 2$	
$X_1 = 0$				
$X_1 = 1$		$\pi_{1\,2}^{(1)(1)}$	$\pi_{1\,2}^{(1)(2)}$	$\pi_1^{(1)}$
		$\pi_{(2)}^{(1)}$	$\pi_{(2)}^{(2)}$	

Both characterizations are equivalent, and the equivalence extends to contingency tables of any dimension.

Limited-information GOF methods disregard information contained in the higher-order marginals of the table. Thus, quadratic forms in, say, univariate and bivariate residuals are used instead of using all marginal residuals up to order n.

GOF Statistics for Assessing Overall Fit

Maydeu-Olivares and Joe (2005, 2006) proposed a family of GOF statistics, M_r, that provides a unified framework for limited information and full information GOF statistics. This family can be written as

$$M_r = N\hat{\mathbf{e}}_r^{'}\hat{\mathbf{C}}\hat{\mathbf{e}}_r, \quad [5]$$

where $\hat{\mathbf{e}}_r$ are the residual proportions up to order r and

$$\mathbf{C} = \boldsymbol{\Gamma}_r^{-1} - \boldsymbol{\Gamma}_r^{-1}\boldsymbol{\Delta}_r(\boldsymbol{\Delta}_r^{'}\boldsymbol{\Gamma}_r^{-1}\boldsymbol{\Delta}_r)^{-1}\boldsymbol{\Delta}_r^{'}\boldsymbol{\Gamma}_r^{-1}. \quad [6]$$

Here, $\boldsymbol{\Gamma}_r$ denotes the asymptotic covariance matrix of the residual proportions up to order r and $\boldsymbol{\Delta}_r$ is a matrix of derivatives of the marginal probabilities up to order r with respect to the model parameters. Two members of this family are, for instance, M_2 and M_n. In M_2 only univariate and bivariate residuals are used. In M_n, all residuals up to order n, the number of variables, are used. When ML estimation is used, M_n is algebraically equal to Pearson's X^2.

The asymptotic distribution of any statistic of the M_r family is chi-square with degrees of freedom (df) = number of residuals used $-q$. For the chi-square approximation to M_r be accurate, the expected frequencies of min $(2r, n)$ marginals need to be large. Thus, for M_n, expected cell frequencies need to be large, but for M_2, where $r = 2$, only expected frequencies for sets of $\min(2r, n) = 4$ variables need to be large (provided $n > 4$). As a result, when only low-order margins are used, the asymptotic p-values are accurate even in gigantic models and small samples. Furthermore, often more power is obtained than when all the information available in the data is used. Consequently, Maydeu-Olivares and Joe suggest testing at the highest level of margins for which a model is identified, discarding higher-order margins. Since most models are identified using only univariate and bivariate information (i.e., they can be estimated using only univariate and bivariate information), M_2 should be the statistic of choice. For instance, the two-parameter logistic item-response theory (IRT) model is identified (it can be estimated) using only univariate and bivariate information. As a result, its fit may be tested using M_2.

GOF Statistics for Piecewise Assessment of Fit

For a piecewise assessment of fit, z-scores for marginal residuals involving one, two, or three variables may be used (i.e., the marginal residuals divided by their *SE*s). It is simpler to extract valuable information from them than from cell residuals. However, when the number of categories is large, there are often too many marginal residuals. In these cases, Pearson's X^2 may be computed for pairs (or if needed, triplets of variables), provided there are enough degrees of freedom for testing. Here, df = number of residuals used–number of estimated parameters involved. However, when used for piecewise assessment of fit, X^2 may yield an undue impression of poor fit (Maydeu-Olivares and Joe, 2006). The use of the M_r statistics in place of X^2 for pairs (or triplets) of variables solves this problem.

GOF Testing with Continuous Observed Data

GOF is a very active area of research in structural-equation modeling (SEM). In classical SEM applications, multivariate models for continuous data (often involving latent variables) are estimated from some summary statistics (typically means and covariances or correlations). For ease of exposition, here we assume that the model is estimated using covariances, but the results can be easily extended to models estimated from other sets of statistics.

Let $\boldsymbol{\sigma}(\boldsymbol{\theta})$ be the $t = n(n+1)/2$ nonredundant population variances and covariances expressed as a function of q model parameters, and let $\mathbf{s}$ be its sample counterpart. The null hypothesis to be tested is $H_0 : \boldsymbol{\sigma} = \boldsymbol{\sigma}(\boldsymbol{\theta})$, that is, that the model holds, against $H_1 : \boldsymbol{\sigma} \neq \boldsymbol{\sigma}(\boldsymbol{\theta})$.

GOF Statistics for Assessing Overall Fit

Two procedures can be used to obtain a GOF statistic. The first procedure is based on using the minimum of the fitted function, $\hat{F}$, multiplied by sample size, N, say $T = N\hat{F}$. This is the usual chi-square test in SEM. The second procedure is based on using a quadratic form in residual summary statistics.

Now, T will only be asymptotically chi-square distributed if the estimator is asymptotically efficient for the distribution of the data. Under multivariate normal assumptions, the efficient estimators are maximum likelihood (ML) and generalized least squares (GLS). Under the asymptotically distribution-free (ADF) distributional assumptions set forth by Browne (1982), the efficient estimator is weighted least squares (WLS). Thus, T is asymptotically distributed as a chi-square with $t-q$ degrees of freedom only when ML or GLS estimation is used under normality, or when WLS estimation is used under ADF assumptions. In all other cases where T is not asymptotically chi-square, its distribution may be approximated by a chi-square if T is scaled by its asymptotic mean or adjusted by its asymptotic mean and variance. These are the so-called Satorra and Bentler (1994) T_s and T_a test statistics, respectively. Thus, if a model is estimated using, for instance, unweighted least squares, T is not asymptotically chi-square, but its distribution can be approximated using a chi-square distribution using Satorra-Bentler corrections.

Regarding the second procedure, Browne (1982) proposed the residual-based statistic

$$T_B = N\hat{\mathbf{e}}'\hat{\mathbf{C}}\hat{\mathbf{e}}, \mathbf{C} = \mathbf{\Gamma}^{-1} - \mathbf{\Gamma}^{-1}\mathbf{\Delta}(\mathbf{\Delta}'\mathbf{\Gamma}^{-1}\mathbf{\Delta})^{-1}\mathbf{\Delta}'\mathbf{\Gamma}^{-1}, \quad [7]$$

where $\hat{\mathbf{e}} = \mathbf{s} - \sigma(\hat{\boldsymbol{\theta}})$, $\mathbf{\Gamma}$ is the asymptotic covariance matrix of the residual covariances, and $\mathbf{\Delta}$ is a matrix of derivatives of the population covariances with respect to the model parameters. $\mathbf{\Gamma}$ may be computed under normality assumptions or under ADF assumptions. This statistic is also asymptotically distributed as a chi-square with $t-q$ degrees of freedom.

Unfortunately, when the data are not normally distributed, it has been repeatedly found in simulation studies that the p-values of Browne's statistic are inaccurate unless the number of variables is small and the sample size is very large. Recently, Yuan and Bentler (1997) have proposed a modification of T_B (with the same asymptotic distribution as T_B) whose p-values are more accurate in small samples and nonnormal data than those for T_B. This is

$$T_{YB} = \frac{T_B}{1 + NT_B/(N-1)^2}. \quad [8]$$

GOF Statistics for Piecewise Assessment of Fit

Z-scores for residuals are used for piecewise assessment of fit. The z-score for the residual covariance between variables i and j is

$$\frac{s_{ij} - \hat{\sigma}_{ij}}{\mathrm{SE}(s_{ij} - \hat{\sigma}_{ij})}, \quad [9]$$

where $\hat{\sigma}_{ij} = \sigma_{ij}(\hat{\boldsymbol{\theta}})$. Browne's and Yuan and Bentler's statistics eqns. [7] and [8] are, in fact, a test based on the joint set of residuals eqn [9].

GOF Indices

For continuous data, the AIC and BIC criteria used for model selection are GOF indices. When ML is not used, then the term $-2L$ in eqn [3] is simply replaced by T, the minimum of the estimated fit function multiplied by sample size, that is

$$\mathrm{AIC} = T + 2q, \ \mathrm{BIC} = T + q\ln(N). \quad [10]$$

The AIC and BIC indices can be computed for any estimator, as no p-value is computed.

In addition, literally dozens of GOF indices have been proposed. Some may be used to assess the overall fit of the model under consideration, whereas others assess the relative fit of the model. A GOF index that may be used to assess the overall fit of a model is the standardized root mean residual (SRMR),

$$\mathrm{SRMR} = \sqrt{\sum_j \sum_{k<j} \left(\frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}} - \frac{\hat{\sigma}_{jk}}{\sqrt{\hat{\sigma}_{jj}}\sqrt{\hat{\sigma}_{kk}}} \right) \Big/ t}. \quad [11]$$

The SRMR may be used to assess the average magnitude of the discrepancies between observed and expected covariances in a correlation metric. Note that there exist slightly different versions of this statistic.

Among GOF indices for relative fit assessment, two popular indices are the Tucker–Lewis index (TLI) and the comparative fit index (CFI), where

$$\mathrm{TLI} = \frac{\frac{T_0}{\mathrm{df}_0} - \frac{T_1}{\mathrm{df}_1}}{\frac{T_0}{\mathrm{df}_0} - 1}, \quad [12]$$

$$\mathrm{CFI} = \frac{(T_0 - \mathrm{df}_0) - (T_1 - \mathrm{df}_1)}{T_0 - \mathrm{df}_0}. \quad [13]$$

Here, M_0 is more restrictive than M_1, the baseline model. Of the two, CFI is normed to lie between 0 and 1, whereas TLI is approximately normed. Almost invariably, they are used to compare the fit of the fitted model against a model that assumes that variables are uncorrelated. When used in this fashion, TLI and CFI values are very large. When comparing a set of theoretically driven models, it may be more interesting to use as a baseline the simplest theoretically driven model under consideration rather than the substantively uninteresting independence model. When used in this fashion, these statistics express in some sort of percentage how much is gained by each of the models under consideration relative to the most parsimonious model. Note that when using the TLI and CFI indices, M_0 need not be a special case of M_1.

Discussion

GOF assessment necessarily involves subjective judgment. Models are just approximations to real-life phenomena. Consequently, any model will be rejected if the sample size is sufficiently large. This should not be taken to imply that GOF testing is meaningless. Rather, it is our view that researchers should always assess the overall GOF of their models using a GOF statistic to assess the magnitude of the discrepancy between the data and the model taking into account sampling variability.

If the selected model fits well, researchers should then:

1. Assess the power of the statistic against meaningful deviations from the selected model, as it may well be that the statistic has no power to distinguish between the selected model and substantively meaningful alternative models.
2. Perform a piecewise assessment of the selected model to examine if, although the model fits well overall,

some parts of the data are not well captured by the model.
3. Consider whether models that cannot be distinguished empirically from the selected model (i.e., equivalent models –see MacCallum *et al.*, 1993), exist.

Researchers should always report substantively interesting models equivalent to their selected model when they are aware of them and argue their choice using substantive arguments, since by definition, a choice between two equivalent models can only be made on substantive, not on empirical grounds.

On the other hand, if the model does not fit well, researchers should:

1. Perform a piecewise assessment of the model attempting to determine the source of the misfit. This may be aided by modification indices (i.e., Lagrange multiplier tests), but see MacCallum *et al.* (1992).
2. Assess the magnitude of the discrepancy between the fitted and expected statistics. For covariance structure analysis, the SRMR provided may be used to examine the average magnitude of the discrepancy. In addition, the magnitude of each standardized discrepancy, that is,

$$\frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}} - \frac{\hat{\sigma}_{jk}}{\sqrt{\hat{\sigma}_{jj}}\sqrt{\hat{\sigma}_{kk}}}, \qquad [14]$$

should be inspected.

In any case, when selecting a model among competing alternatives, they should strive for model parsimony. To this end, they may inspect GOF indices that penalize adding parameters by, such as the AIC or BIC indices to the model. Alternatively, they may compute the CFI index, using as a baseline, the most parsimonious substantively meaningful model considered.

Finally, when many variables are modeled, it is unrealistic to expect that any parsimonious model will fit well: the overall test statistics will be large, because some parts of the data will not be well reproduced by any parsimonious model. In this context, Browne and Cudeck's (1993) proposal of assessing whether a model fits the data closely, is an attractive alternative to assessing whether the model fits exactly. Consider the root mean square error of approximation (RMSEA) index,

$$\text{RMSEA} = \sqrt{\max\left(\frac{T - \text{df}}{N \times \text{df}}, 0\right)}. \qquad [15]$$

Like other indices, the RMSEA penalizes models with too many parameters. Unlike the AIC or BIC criteria, the RSMEA is bounded below by 0. Furthermore, Browne and Cudeck derived its asymptotic distribution. Testing $H_0 : \text{RMSEA} = 0$ is equivalent to testing $H_0 : \boldsymbol{\sigma} = \boldsymbol{\sigma}(\boldsymbol{\theta})$. For large models, this null hypothesis of exact fit may be too stringent, and Browne and Cudeck suggested testing instead, $H_0 : \text{RMSEA} \leq 0.05$. This is a test of close fit, where close is arbitrarily defined as $\text{RMSEA} \leq 0.05$. Moreover, model selection may be aided by inspecting the confidence intervals around their respective RMSEAs (Steiger, 2007).

Numerical Examples

Discrete Data

We fitted a two-parameter logistic model with a normally distributed latent trait to the Law school admission test (LSAT) 7 data. A thousand observations are available on five binary variables. The model was estimated by ML. **Table 1** gives some relevant GOF statistics. As these data are not sparse, the p-values for X^2 and G^2 are accurate, and they are similar to that of the new test M_2, which yields accurate p-values even for highly sparse data.

Using a 5%-significance level, X^2 and G^2 suggest that the model be barely accepted, whereas M_2 suggests that the model be barely rejected. Statistics with a higher value to degrees-of-freedom ratio are generally more powerful, and the results of **Table 1** suggest that M_2 has higher power.

The cell residuals are not very helpful for piecewise assessment of fit. Significant residuals (at the 5% level) are obtained for patterns (10000), (01001), and (01000), which might suggest that the fit of the model could improve by dropping Item 2. The inspection of univariate and bivariate residuals offers a different picture. Significant residuals are obtained for (1,3), (1,5), (1,4), and (2,3), which suggest that the model misfits because of item 1. This is indeed the case as reflected in **Table 2**, where we provide Pearson's X^2 statistics after dropping one item at a time.

Continuous Data

We model the responses of 438 US respondents to the five items of the satisfaction with life scale (SWLS). The items

Table 1 Goodness-of-fit (GOF) results for the LSAT 7 data

Stat	*Value*	*df*	*p-value*	*Value/df*
X^2	32.48	21	0.05	1.55
G^2	31.70	21	0.06	1.51
M_2	11.94	5	0.04	2.39

Table 2 GOF results for the LSAT 7 data dropping one item at a time, 7 df

Item dropped	*1*	*2*	*3*	*4*	*5*
X^2	5.01	9.52	8.59	18.68	9.86
p-value	0.66	0.22	0.22	0.01	0.20

Table 3 GOF results for the SWLS data

NT				*ADF*			
Stat	*Value*	*df*	*p*	*Stat*	*Value*	*df*	*p*
T	10.38	3	0.02	T_B	8.72	3	0.03
T_B	10.38	3	0.02	T_{YB}	8.55	3	0.04
				T_S	7.17	3	0.04
				T_a	6.90	2.89	0.07

are rating scales with seven response alternatives and will be treated as continuous. A two-factor model where "In most ways my life is close to my ideal," "The conditions of my life are excellent," and "I am satisfied with my life," are taken as indicators of the factor satisfaction with present life, the and "I am satisfied with my life," "So far I have gotten the important things I want in life," and "If I could live my life over, I would change almost nothing," are taken as indicators of the factor of satisfaction with past life. The factors are correlated. **Table 3** lists a number of GOF statistics obtained using the maximum-likelihood fitting function either under normality assumptions (NT) or under asymptotically distribution-free (ADF) assumptions.

Under NT, $T = N\hat{F}$, the minimum of the ML fit function multiplied by the sample size is provided, along with Browne's test (eqn [7], T_B. Under ADF assumptions, **Table 3** lists Browne's test, Yuan and Bentler's test (eqn [8], T_{YB}, and the Satorra-Bentler mean and mean and variance adjustments to T, T_s, and T_a, respectively. Notice that for T_a, the degrees of freedom are estimated as real numbers.

The model does not fit very well, and there is not much difference between the results under NT or ADF assumptions. All statistics yield similar p-values. Inspection of the z-scores for the residual covariances reveals significant residual covariances among items from different factors. The magnitude of the residuals is not large, however. The average standardized residual is SRMSR = 0.015, and the largest standardized residual is 0.04. It appears that the model yields a close enough fit. Indeed, the RMSEA (eqn. [15] obtained is 0.056, and the p-value for testing whether the population RMSEA is smaller than 0.05, is 0.35.

Concluding Remarks

Our presentation has focused on models where the data can be summarized using some statistics (proportions, covariances, etc.). The fit of many interesting models cannot be assessed using summary statistics. For instance, the fit of a linear regression model cannot be assessed using covariances (i.e., within a SEM framework) because there are zero degrees of freedom. In the context of linear regression and related models, R^2 is sometimes described as a GOF statistic. However, R^2 is actually a coefficient of determination, the proportion of the dependent variable that can be predicted from the independent variables. A linear regression model can fit the data perfectly, yet, R^2 will be zero if the slope is zero. Pure GOF statistics exist in regression and related models only in the presence of replicates (i.e., repeated observations for the same level of the predictors). In the general linear model, they are generally referred to as lack-of-fit tests. When no replicates exist, then, the observations must be grouped in some way to assess the GOF of the model. A typical example is the Hosmer–Lemeshow GOF statistic for logistic regression (see Hosmer and Lemeshow, 2000).

Clearly, GOF assessment has been more extensively developed in SEM than in other areas. New developments in GOF assessment for multivariate discrete data are strongly related to SEM procedures, and we expect further developments in GOF assessment procedures for multivariate discrete data along the lines of SEM developments. For instance, the notion of testing whether a model fits closely the data (as opposed to exactly), is yet to be brought into the multivariate discrete arena. More research is also needed on GOF assessment when the observed dependent variables are of mixed type (continuous and categorical). Finally, further research is needed on GOF-assessment procedures when data arises from complex sampling schemes, such as those found in multilevel modeling.

In closing, model fit (i.e., absolute GOF) is no guarantee of a model's usefulness. A model may reproduce the data at hand well and yet be useless for the purpose it was developed. On the other hand, a model may fit poorly and yet yield useful predictions. In this context, the fact that a model fits poorly simply means that in principle, a model could be found to reproduce the data better, whose predictions could be very different.

See also: Educational Data Modeling; Item Response Theory; Latent Class Models; Model Selection; Multivariate Linear Regression; Structural Equation Models.

Bibliography

Browne, M. W. (1982). Covariance structures. In Hawkins, D. M. (ed.) *Topics in Applied Multivariate Analysis*, pp 72–141. Cambridge: Cambridge University Press.

Browne, M. W. and Cudeck, R. (1993). Alternative ways of assessing model fit. In Bollen, K. A. and Long, J. S. (eds.) *Testing Structural Equation Models*, pp 136–162. Newbury Park, CA: Sage.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. New York: Wiley.

MacCallum, R. C., Roznowski, M., and Necowitz, L. B. (1992). Model modification in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin* **111**, 490–504.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., and Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin* **114**, 185–199.

Maydeu-olivares, A. and Cai, L. (2006). A cautionary note on using G^2 (dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research* **41**, 55–64.

Maydeu-olivares, A. and Joe, H. (2005). Limited and full information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association* **100**, 1009–1020.

Maydeu-Olivares, A. and Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika* **71**, 713–732.

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

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modelling. *Personality and Individual Differences* **42**, 893–898.

Tollenaar, N. and Mooijart, A. (2003). Type I errors and power of the parametric goodness-of-fit test. Full and limited information. *British Journal of Mathematical and Statistical Psychology* **56**, 271–288.

Yuan, K.-H. and Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association* **92**, 767–774.

Yuan, K.-H. and Bentler, P. M. (2004). On chi-square difference and *z* tests in mean and covariance structure analysis when the base model is mispecified. *Educational and Psychological Measurement* **64**, 737–757.

Further Reading

Bartholomew, D. J. and Tzamourani, P. (1999). The goodness of fit of latent trait models in attitude measurement. *Sociological Methods and Research* **27**, 525–546.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin* **107**, 238–246.

Bollen, K. A. and Stine, R. (1992). Bootstrapping goodness of fit measures in structural equation models. *Sociological Methods and Research* **21**, 205–229.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology* **44**, 108–132.

Browne, M. W., MacCallum, R. C., Kim, C.-T., Anderson, B., and Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods* **7**, 403–421.

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

Joe, H. and Maydeu-Olivares, A. (2006). On the asymptotic distribution of Pearson's X^2 in cross-validation samples. *Psychometrika* **71**, 587–592.

MacCallum, R. C., Browne, M. W., and Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods* **11**, 19–35.

MacCallum, R. C., Browne, M. W., and Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods* **1**, 130–149.

Mavridis, D., Moustaki, I., and Knott, M. (2007). Goodness-of-fit measures for latent variable models for binary data. In Lee, S.-Y. (ed.) *Handbook of Latent Variables and Related Models*, pp 135–162. Amsterdam: Elsevier.

McDonald, R. P. and Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods* **7**, 64–82.

Satorra, A. and Saris, W. E. (1993). *Power Evaluations in Structural Equation Models*, pp 181–204. Newberry Park, CA: Sage.

Swaminathan, H., Hambleton, R. K., and Rogers, H. J. (2007). Assessing the fit of item response models. In Rao, C. R. and Sinharay, S. (eds.) *Handbook of Statistics, Vol. 26: Psychometrics*, pp 683–718. Amsterdam: Elsevier.

Yuan, K. H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research* **40**, 115–148.