4 Structural Equation Models With Nonnormal Variables

Problems and Remedies

STEPHEN G. WEST

JOHN F. FINCH

PATRICK J. CURRAN

Over the past 15 years, the use of structural equation modeling has become increasingly common in the social and behavioral sciences. Enthusiastic recognition by researchers of the advantages of the structural equation modeling approach and an eagerness to implement this potentially powerful methodology has also brought with it inappropriate use of the technique. One major source of inappropriate usage has been the failure of investigators to satisfy the scaling and normality assumptions upon which estimation and testing are based. The commonly used approaches to estimating the parameters of structural equation models, maximum likelihood and normal theory generalized least squares, assume that the measured variables are continuous and have a multivariate normal distribution. In practice, current applications of the structural equation modeling approach to real data often involve violations of these assumptions.

AUTHORS' NOTE: S. G. West was supported by NIMH grant P50MH39246 during the writing of this chapter. We thank Leona Aiken, William L. Cook, William R. Shadish, Jr., and Rick Hoyle for their comments on an earlier version of this chapter.

In some substantive areas, the measured variables used by researchers are dichotomous or ordered categories (e.g., "agree," "no preference," "disagree") rather than truly continuous. In other areas, the measured variables are continuous but their distributions depart dramatically from normality (e.g., measures of amount of substance use). Micceri (1989) analyzed over 400 large data sets, finding that the great majority of data collected in behavioral research do *not* follow univariate normal distributions, let alone a multivariate normal distribution. Yet researchers often ignore these assumptions. For example, Breckler (1990) identified 72 articles in personality and social psychology journals that had used structural equation modeling and found that only 19% acknowledged the normal theory assumptions, and fewer than 10% explicitly considered whether these assumptions had been violated.

Given that real data often fail to satisfy the underlying scaling and normality assumptions, there has been growing interest in determining the robustness of structural equation modeling techniques to violations of the scaling and normality assumptions and in developing alternative remedial strategies when these assumptions are seriously violated. These topics are the focus of the present chapter.

Overview of Normal Theory Estimation

As discussed in Chapters 1 and 3, the objective of estimation is to minimize the magnitude of the set of differences between each element in S and the corresponding element in $\Sigma(\hat{\theta})$. Recall that S is the sample covariance matrix calculated from the observed data and $\Sigma(\hat{\theta})$ is the covariance matrix implied by a set of parameter estimates $\hat{\theta}$ for the hypothesized model. Throughout the presentation below, all parameters that are estimated will be grouped in a vector θ.

The two most commonly used estimation techniques are maximum likelihood (ML) and normal theory generalized least squares (GLS). Both techniques are based on the same set of assumptions, yield very similar estimates, and have the same desirable statistical properties. These techniques are discussed in more detail in Chapter 3 and by Bollen (1989b); here we briefly review the assumptions and properties of the GLS estimator to set the stage for our later discussion of nonnormality.

The generalized least squares fitting function, F_{GLS}, can be expressed as

$$F_{GLS} = \tfrac{1}{2}\mathrm{tr}\left[([S - \Sigma(\hat{\theta})]W^{-1})^2\right]. \quad (4.1)$$

In this equation, S represents the observed covariance matrix, $\Sigma(\hat{\theta})$ represents the covariance matrix implied by the hypothesized model, W^{-1} represents a weight matrix, and "tr" is the trace operator, which takes the sum of the elements on the main diagonal of the matrix, here the matrix resulting from the operations within the large brackets. Minimization of this fitting function involves minimization of the weighted squared discrepancies between S and $\Sigma(\hat{\theta})$. Like other members of the class of weighted least squares procedures, GLS requires the selection of the weight matrix. The most common choice for W^{-1} is S^{-1}, which weights the squared discrepancies between S and $\Sigma(\hat{\theta})$ according to their variances and covariances with other elements. This choice is based on two assumptions. First, $E(s_{ij})$, the expected value of the sample covariance between x_i and x_j, is assumed to equal σ_{ij}, the corresponding covariance in the population. Second, the large sample distribution of the elements of S is assumed to be multivariate normal. If these assumptions are satisfied, GLS estimates have several desirable statistical properties.

1. The parameter estimates are asymptotically unbiased: On average, in large samples, they neither overestimate nor underestimate the corresponding population parameter (i.e., $E[\hat{\theta}] = \theta$, where $E[\hat{\theta}]$ is the expected value of the estimate of θ).
2. The parameter estimates are consistent: They converge in probability to the true value of the population parameter being estimated as sample size increases.
3. The parameter estimates are asymptotically efficient: With increasing N, they have minimum variance.
4. $(N - 1)F_{GLS}$ approximates a chi-square distribution in large samples, permitting tests of the fit of the model to the data.

Recall, however, that these desirable statistical properties of the GLS estimator (and the ML estimator; see Bollen, 1989b) are contingent on meeting several assumptions. These assumptions include that a very large (asymptotic) sample size is employed, the observed variables are continuous, the measured variables have a multivariate normal distribution, and the model estimated is a valid one. When these assumptions are *not* met, there is no guarantee in statistical theory that the desirable

properties will continue to hold. Consequently, the robustness of the estimators to violations of assumptions becomes an important issue for empirical study.

Effects and Detection of Nonnormality

THEORETICAL BASIS FOR THE PROBLEM

Potential problems in estimation of structural equation models are introduced when the distribution of the observed variables departs substantially from multivariate normality. As can be seen from Equation 4.1, the parameter estimates are derived from information in S, the sample covariance matrix, and W^{-1}, the optimal weight matrix. When the observed variables are (a) continuous but nonnormal, (b) dichotomous, or (c) ordered categories, the information in S or W^{-1} or both may be incorrect. As a result, estimates based on S and W^{-1} may also be incorrect.

Continuous, Nonnormal Variables. As we saw in the discussion of estimation, the variation in the measured variables is completely summarized by the sample covariances only when multivariate normality is present. If multivariate normality is violated, the variation of the measured variables will not be completely summarized by the sample covariances; information from higher-order moments is needed. In this situation, S^{-1} is no longer the correct estimator of W^{-1}. The parameter estimates do remain unbiased and consistent (i.e., as sample size grows larger, $\hat{\theta}$ converges to θ), but they are no longer efficient. These results suggest that theoretically two important problems will occur with normal theory estimators (ML, GLS) when the observed variables do not have a multivariate normal distribution. (a) The χ^2 goodness-of-fit test is not expected to produce an accurate assessment of fit, rejecting too many (> 5%) true models. (b) Tests of all parameter estimates are expected to be biased, yielding too many significant results.

Coarsely Categorized Variables. Investigations of the effects of coarse categorization of continuous variables (e.g., Bollen & Barb, 1981) have found that the Pearson correlation coefficient between two continuous variables is generally higher in magnitude than the correlation between the same variables when they have been divided up into a

set of ordered categories. The greatest attenuation occurs when few categories are employed (i.e., fewer than five) for either variable involved in the correlation and when the categorized variables are skewed, particularly in opposite directions. These findings imply that coarse categorization of continuous variables can theoretically be expected to lead to biased χ^2 tests of model fit, parameter estimates, standard errors, and tests of parameter estimates.

DETECTING DEPARTURES FROM NORMALITY

Skewness and Kurtosis, Univariate and Multivariate. A number of procedures are available for assessing the univariate and multivariate normality of the measured variables. These procedures depend on the calculation of higher order moments: A moment is defined as $(1/N)\sum(x - \mu)^k$, where N is sample size, x is an observed score, μ is the population mean, and k is the order of the moment ($k = 1$ for the first-order moment; $k = 2$ for the second-order moment, etc.). When univariate normality is satisfied, only the first- and second-order moments (mean and variance) are needed to describe fully the distribution of the measured variables—the standardized third-order moment is 0 and the standardized fourth-order moment is technically 3 for a normal distribution. Univariate distributions that deviate from normality, however, possess significant nonzero skewness and kurtosis that are reflected in the standardized third- and fourth-order moments, respectively. Nonzero skewness is indicative of a departure from symmetry. Negative skewness indicates a distribution with an elongated left-hand tail; positive skewness indicates a distribution with an elongated right-hand tail (relative to the symmetrical normal distribution). Kurtosis, which is particularly important for statistical inference, indicates the extent to which the height of the curve (probability density) differs from that of the normal curve. Positive kurtosis is associated with distributions with long, thin tails, whereas negative kurtosis is associated with shorter, fatter tails relative to the normal curve. To simplify interpretation, many computer packages subtract 3 from the standardized fourth-order moment so that kurtosis will be 0 for a normal curve. We follow this convention in reporting values of kurtosis in this chapter.

Examinations of the skewness and kurtosis of the univariate distributions provide only an initial check on multivariate normality. If any of the observed variables deviate substantially from univariate normality, then the multivariate distribution cannot be multinormal. However, the converse is *not* true: Theoretically, all of the univariate distributions may be normal, yet the joint distribution may be substantially multivariately nonnormal. Consequently, it is also important to examine multivariate measures of skewness and kurtosis developed by Mardia (1970; see also D'Agostino, 1986).

The Mardia measures construct functions of the third- and fourth-order moments, which possess approximate standard normal distributions, thereby permitting tests of multivariate skewness and multivariate kurtosis. The Mardia measure of multivariate kurtosis, which is particularly important for structural equation modeling (Browne, 1982), is available in the EQS (Bentler, 1992a) and PRELIS (Jöreskog & Sörbom, 1993c) computer software packages.

Outliers. Outliers are extreme data points that may affect the results of structural equation modeling, even when the remainder of the data are well distributed. Outliers typically occur because of errors in responding by subjects or data recording errors, or because a few respondents may represent a different population from the target population under study. Outliers can potentially have dramatic effects on the indices of model fit, parameter estimates, and standard errors. They can also potentially cause improper solutions, in which estimates of parameters are outside the range of acceptable values (e.g., Heywood cases in which estimates of error variance are < 0; see Dillon, Kumar, & Mulani, 1987). Possible corrective actions for outliers include checking and correction of the data for the extreme case, dropping the extreme case, redefinition of the population of interest, or respecification of the model, with the appropriate remedy depending on the apparent source of the outlier.

Two general approaches can be used to detect outliers in the context of structural equation models. The first, a model-independent approach, is to identify any deviant cases whose values diverge sharply from the mass of data points. Univariately, this can be accomplished by visual examination of the plots of each measured variable, identifying cases that are several standard deviations from the mean of the distribution and not close to other observations. Multivariately, leverage statistics, such as Mahalanobis distance available in major regression diagnostic packages, identify extreme points in multivariate space (see Chatterjee & Yilmaz, 1992). Alternatively, Bentler (1989) has proposed identify-

ing the cases that have the greatest contribution to Mardia's measure of multivariate kurtosis. Typically, all measured variables would be considered together in these analyses.

The second approach is to identify observed data points that are extreme relative to their predicted value based on a specific model. Bollen and Arminger (1991) have proposed a method based on factor scores, which represent each case's predicted score on the hypothetical factor. These factor scores, in turn, are used to estimate a set of predicted scores on the measured variables for each case. Raw residuals representing the difference between the predicted and the observed scores for each case on each measured variable are calculated. The residuals are standardized ($M = 0$; SD = 1), using procedures described in Bollen and Arminger (1991), and then plotted and visually examined to detect possible outliers.

RESULTS OF EMPIRICAL STUDIES OF NONNORMALITY

Continuous, Nonnormal Variables. Several simulation studies have assessed the performance of the normal theory ML and GLS estimators for a variety of CFA models under diverse conditions of nonnormality and sample size (Browne, 1984a; Curran, West, & Finch, 1994; Finch, Curran, & West, 1994; Hu, Bentler, & Kano, 1992). In these studies, the value of each parameter is set to a known value in the population. This value is then compared with the mean of a large number of empirical estimates to study the effects of specified levels of nonnormality. The following conclusions have been reached:

1. ML and GLS estimators produce χ^2 values that become too large when the data become increasingly nonnormal. For example, Curran et al. (1994) investigated a three-factor, nine-indicator confirmatory factor analysis model in which each measured variable was highly nonnormal (skewness = 3; kurtosis = 21). Compared to the expected χ^2 of 24, the mean of χ^2 from 200 simulations was 37.4 (approximate 50% overestimate) when sample size was 1000 in each simulation. Compared to the expected Type 1 error rate of 5%, 48% of the true models in the population were rejected under these conditions.

2. The GLS and particularly the ML estimator produce χ^2 values that are slightly too large when sample sizes are small, even when multivariate normality is present. For example, in the Curran et al. (1994) study, when the sample size was 50 and the observed variables were multivariate normal, the mean χ^2 of 200 simulations was 26.7 (10% overestimate) and 12% of the true models in the population were rejected. Simulations by Anderson and Gerbing (1984) and Boomsma (1983) have also found that decreasing sample size and increasing nonnormality lead to increases in the proportion of analyses that fail to converge or that result in an improper solution (Heywood case).

3. Nonnormality leads to modest underestimation of fit indexes such as the Normed Fit Index (NFI; Bentler & Bonett, 1980), the Tucker and Lewis (1973) Index (TLI), and the Comparative Fit Index (CFI; Bentler, 1990). (See Tanaka, 1993, for an overview of fit indexes.) For example, Curran et al. (1994) found that when using maximum likelihood estimation with a sample size of 100, the mean CFI for a correctly specified model was .97 (3% underestimate), compared to the expected value of 1.00 when each of the measured variables was highly nonnormal (skewness = 3; kurtosis = 21). The TLI and the CFI are modestly underestimated, whereas the NFI is severely underestimated at low sample sizes (e.g., mean NFI = .81 vs. 1.00 expected at $N = 50$ under multivariate normality; see also Marsh, Balla, & McDonald, 1988).

4. Nonnormality leads to moderate to severe underestimation of standard errors of parameter estimates. For example, Finch et al. (1994) studied the standard errors of parameter estimates in confirmatory factor analysis models. When the measured variables were highly nonnormal (skewness = 3; kurtosis = 21), the standard errors of correlations between factors (ϕ) were underestimated by about 25%, whereas the standard errors of factor loadings (λ) and the specific factors (error variances; θ) were underestimated by approximately 50%. Such substantial underestimates in standard errors imply that tests of parameter estimates will not be trustworthy under conditions of nonnormality.

Coarsely Categorized Variables. Several simulation studies (Babakus, Ferguson, & Jöreskog, 1987; Boomsma, 1983; Muthén & Kaplan, 1985) have evaluated the performance of the normal theory ML and GLS estimators when continuous normally distributed measured variables are divided into ordered categories. Once again, a variety of CFA models and rules for categorizing the continuous variables have been utilized. These studies have led to the following conclusions:

1. The number of categories per se has relatively little impact on the χ^2 goodness-of-fit test when the distribution of the categorized variables is approximately normal. As the distributions of the categorized

variables become increasingly and particularly differentially skewed (e.g., variables skewed in opposite directions), the χ^2 values become inflated.

2. Factor loadings and factor correlations are only modestly underestimated as long as the distribution of the categorized variables is approximately normal. However, underestimation becomes increasingly serious as (a) there are fewer categories (e.g., two or three), (b) the magnitude of skewness increases (e.g., > 1), and (c) there is a differential degree of skewness across variables.

3. Estimates of error variances (specific factors) are more severely biased than other parameter estimates by each of the influences noted under (2). Relatedly, correlations may be spuriously obtained between the error variances associated with items having similar degrees of skewness. When there are only a small number (e.g., two) of categories, the degree of skewness is determined by the percentage of subjects in the study agreeing with (or passing) the item. Thus a set of items with similar agreement rates (e.g., 15% to 20%) can give rise to a spurious factor (so-called "difficulty factor") reflecting only the common degree of skewness among the items.

4. Estimated standard errors for all parameters are too low, particularly when the distributions are highly and differentially skewed. This means that tests of parameter estimates may not be trustworthy.

Remedies for Multivariate Nonnormality

ALTERNATIVE ESTIMATION TECHNIQUES

As we saw above, the problem of nonnormality can arise in two different contexts: poorly distributed continuous variables or coarsely categorized continuous variables. Estimation-based remedies to these two problems differ. However, these techniques share the common goal of yielding χ^2 tests and estimates of standard errors that more closely approximate their true values.

The Asymptotically Distribution Free Estimator. Browne (1984a) developed an alternative estimator that does not assume multivariate normality of the measured variables. His "asymptotically distribution free" (ADF) estimation procedure is based on the computation of a general weight matrix, W, and GLS estimation. The key to ADF estimation is the use of an optimal weight matrix that consists of a combination of second- and fourth-order terms. W is a covariance matrix of the elements in S, which contains both variances and covariances. Thus the ADF weight matrix has many more elements than the normal theory GLS weight matrix (S^{-1}); however, it has the desirable property of simplifying to the normal theory matrix (S^{-1}) under conditions of multivariate normality (i.e., fourth-order moments = 0). Because of the link to the normal theory GLS fitting function, the ADF estimator is sometimes referred to as the arbitrary generalized least squares (AGLS) estimator.

The ADF estimator produces asymptotically (large sample) unbiased estimates of the χ^2 goodness-of-fit test, parameter estimates, and standard errors. These are major theoretical advantages relative to the normal theory-based ML and GLS estimators, which, as was shown above, produce biased test statistics and standard errors under conditions of multivariate nonnormality. However, the ADF estimator is associated with two important practical limitations. First, the ADF estimator is computationally demanding. The calculation of the ADF fitting function requires the inversion of the ADF optimal weight matrix. In CFA with p measured variables, W is a $p^* \times p^*$ matrix, where p^* is $\frac{1}{2}p(p+1)$, the number of unique elements in S. For example, with 15 measured variables it is necessary to invert a 120 by 120 weight matrix consisting of 14,400 unique elements. With more than 20 to 25 measured variables, implementation of the methodology becomes impractical, even given modern high speed computers (Bentler, 1989). Second, the calculation of the matrix of fourth-order moments requires a large sample size to produce stable estimates (Jöreskog & Sörbom, 1992). This sample-size based limitation is a serious one, as we will see below.

SCALED χ^2 Statistic and Robust Standard Errors. Although the normal theory χ^2 statistic does not follow the expected χ^2 distribution under conditions of nonnormality, it can be corrected or rescaled to approximate the referenced χ^2 distribution. Satorra and Bentler (see Satorra, 1990) have developed the statistical theory underlying this rescaling. The normal theory χ^2 (from ML or GLS) is divided by a constant k, whose value is a function of the model-implied residual weight matrix, the observed multivariate kurtosis, and the degrees of freedom for the model. As the degree of multivariate kurtosis increases, so does k, subsequently leading to a greater downward adjustment of

the normal theory χ^2. The same theory underlying the SCALED χ^2 statistic can also be applied to the computation of robust standard errors. These standard errors can theoretically be considered to be adjusted for the degree of multivariate kurtosis. The SCALED χ^2 and robust standard errors are available in the EQS program.

Bootstrapping. Modern, computationally intensive statistical methods provide a completely different approach to tests of goodness-of-fit and parameter estimates. Rather than relying on the theoretical distributions of classical test statistics (e.g., χ^2, normal), we can imagine taking repeated samples from a population of interest. For each sample, we calculate the parameter estimates of interest resulting in an *empirical* sampling distribution. In cases in which the assumptions of the classical test statistics are severely violated, the empirical distribution that describes the actual distribution of the estimates from this population will be substantially more accurate than the theoretical distribution.

Efron and his coworkers (e.g., Efron & Tibshirani, 1986; see Mooney & Duval, 1993) have shown that the empirical sampling distribution can often be reasonably approximated based on data from a single sample. In the bootstrapping procedure, repeated samples of the same size are taken from the original sample *with replacement* after each case is drawn. To illustrate, imagine that the original sample consists of cases (1, 2, 3, 4). Three possible bootstrap samples from this original sample are (1, 4, 1, 1), (2, 3, 1, 3), and (4, 2, 2, 4). Note that the elements can be repeated in the bootstrap samples and that they are of the same size as the original sample. By taking a large number of bootstrap samples from the original sample, the mean and variance of the empirical bootstrap sampling distribution can be determined.

The bootstrap approach is simple conceptually and computationally, given the increasing availability of software to implement bootstrap resampling, including some of the structural equation modeling packages. Two related complexities arise in application. First, as Bollen and Stine (1992, p. 207) emphasize: "The success of the bootstrap depends on the sampling behavior of a statistic being the same when the samples are drawn from the empirical distribution and when they are taken from the original population." Conditions under which this assumption appears to hold are discussed in Efron and Tibshirani (1986). Second, the bootstrap is often more usefully applied to understand a portion or a transformation of the statistic of interest. Bollen and Stine (1992) have shown that the simple bootstrap approach to the χ^2 goodness-of-fit test

for a properly specified model in CFA often produces inaccurate results under conditions of multivariate normality. Even with a properly specified model in the population, the original sample will reflect some sampling fluctuation (e.g., s_{ij} in the sample will not, in general, equal σ_{ij}). The expected value of the χ^2 for the set of bootstrap samples constructed from the original sample will typically *not* be equal to the expected value of the χ^2 (i.e., the df for the model) for a set of samples taken from the population. Consequently, the bootstrap distribution will follow a noncentral χ^2 distribution (which reflects the fluctuation present in the original sample), rather than the usual central χ^2 distribution specified by statistical theory. Bollen and Stine (1992) present a transformation that is a complex function of the original data in the sample and its covariance matrix that minimizes this problem. Evaluations of Bollen and Stine's approach have also shown reasonable performance compared to the values expected from statistical theory for the χ^2 test statistic and the standard errors of direct effects and indirect effects under conditions of multivariate normality.

Empirical Studies of Alternative Estimation Procedures. A number of simulation studies have examined the performance of the ADF estimator, the SCALED χ^2 statistic and robust standard errors, or both (Chou & Bentler in Chapter 3, this volume; Chou, Bentler, & Satorra, 1991; Curran et al., 1994; Finch et al., 1994; Hu et al., 1992; Muthén & Kaplan, 1985, 1992). To date, no large simulation studies have investigated the performance of the bootstrapping approach with diverse nonnormal distributions. The following conclusions may be reached about the ADF and rescaling approaches:

1. All studies have found that the ADF procedure produces χ^2 statistics that are far too high when sample sizes are small to moderate. For example, in the Curran et al. (1994) study, in which the expected χ^2 was 24, when the sample size was 100, the ADF-based χ^2 was 36.4 (50% overestimate) when the distribution was multivariate normal and 44.8 (approximate 90% overestimate) when all measured variables were highly nonnormal (skewness = 3; kurtosis = 21). In contrast, the corresponding values of the SCALED χ^2 statistics were 25.2 (5% overestimate) and 26.8 (10% overestimate), respectively. Under these conditions, the ADF estimator rejected 68% of models that were true in the population, whereas the SCALED χ^2 statistic rejected only 10% of models that were true in the population.

All studies have shown that very large samples are required for adequate performance of the ADF-based χ^2 statistic. Sample sizes of 1000 appear to be necessary with relatively simple models under typical conditions of nonnormality (Curran et al., 1994). Perhaps 5000 cases are necessary for more complex models, less favorable nonnormal conditions, or both (Hu et al., 1992). The SCALED χ^2 statistic appeared to provide good estimates of χ^2 for samples of size 200 and higher.

2. Finch et al. (1994) found that when sample size was 100 and the data were highly nonnormal (skewness = 3; kurtosis = 21), the ADF estimates of the standard error underestimated the empirical standard errors of the factor correlations by 25% and the standard errors of the factor loadings and error variances (specific factors) by approximately 35%. The performance of the Satorra-Bentler robust standard errors was only modestly better under these conditions, with the standard errors being underestimated by approximately 20% for the factor correlations and 25% for the factor loadings and specific factors. The robust standard errors provided generally accurate estimates beginning at a sample size of 200 for moderately nonnormal (skewness = 2; kurtosis = 7) and 500 for highly nonnormal observed variables.

Coarsely Categorized Variables. As we saw earlier, coarse categorization of continuous variables produces bias not only in the χ^2 test-of-fit and standard errors of parameter estimates, but also in the parameter estimates themselves. Muthén (1984) has developed an alternative estimator, which he termed the CVM (for continuous/categorical variable methodology) estimator. The CVM estimator permits the analysis of any combination of dichotomous, ordered polytomous, and interval-scaled measured variables. Unlike traditional normal theory methods, the CVM estimator can yield unbiased, consistent, and efficient parameter estimates when observed variables are dichotomous or ordered categories.

The CVM approach to estimation is based on a strong assumption: A continuous normally distributed ($M = 0$, $\sigma^2 = 1.0$) latent response variable, y^*, is assumed to underlie each measured variable, y. For dichotomous variables, a response of "yes" would be observed if the individual's standing on the underlying normally distributed y^* dimension is greater than a threshold value. A response of "no" would be observed if the individual's standing was below the threshold. Generalizing to ordered categorical variables, the observed response category is assumed to depend on the individual's standing on the normally distributed underlying y^* variable, relative to a set of response thresholds. In the case of a continuous measured variable, y and y^* are assumed to be equivalent.

Because the categorical and/or nonnormally distributed y variables are assumed to be only approximations of the underlying normally distributed y^*s, a distinction is drawn between the covariance structure of the ys and the covariance structure of the underlying y^*s. When one or more observed variables are categorical, the covariance structure of the ys will differ from the covariance structure of the y^*s in important respects. In general, measures of association between categorical variables will be attenuated relative to the underlying, continuous y^*s. A solution in this case is to calculate measures of association between the y^*s based on tetrachoric, polychoric, and polyserial correlations between the measured y variables. The objective of the CVM approach, then, is to reproduce this estimated covariance structure of the y^* variables.

Note that this approach will be theoretically reasonable only in some cases. For example, for many attitude items, the researcher will be more interested in the relationships among the normally distributed, continuous underlying latent variables than in the simple relationships between the observed "agree" versus "disagree" responses on the items. For other continuously distributed variables such as current drug use ("yes" vs. "no"), it is difficult to conceive of a *normally distributed* underlying latent variable. Finally, some variables such as gender are inherently categorical, so no continuous underlying variable could exist.

The CVM approach once again utilizes a weighted least squares estimator (Muthén, 1984). The fitting function minimized by this estimator is of the form

$$F_{\text{WLS}} = [S - \sigma(\hat{\theta})]' \, W^{-1} \, [S - \sigma(\hat{\theta})], \tag{4.2}$$

where p is the number of measured variables, S is a $p^* \times 1$ vector containing the nonredundant elements of the sample covariance matrix, $\sigma(\hat{\theta})$ is the corresponding $p^* \times 1$ vector from the model implied covariance matrix $\Sigma(\hat{\theta})$, and W^{-1} is a $p^* \times p^*$ weight matrix. Here p^* is defined as $\frac{1}{2}p(p+1)$. When S contains Pearson correlations (or covariances) for normally distributed interval scaled measured variables, the fitting function simplifies to the normal theory GLS estimator discussed previously. Muthén's CVM approach is very general and can be applied to ordered categories through the use of polychoric correlations and con-

tinuous variables that have been censored or truncated through the use of tobit correlations (Muthén, 1991). Combinations of these types of variables can also be addressed.

Muthén's CVM approach also has some significant limitations. Like the ADF estimator, the estimation of the weight matrix places severe practical limits on the number of variables that can be considered (maximum is about 25). The use of the CVM estimator also requires that large samples be used (at least 500-1000 cases, depending on the complexity of the model). Nonetheless, simulation studies to date (see, e.g., Muthén & Kaplan, 1985; Schoenberg & Arminger, 1989) have shown good performance of the CVM estimator relative to ML, GLS, and ADF estimators. The differences in performance are most apparent under the conditions identified above when ML and GLS perform poorly: The observed variables have a small number (two to three) of categories and are highly (> 1 in magnitude) and differentially skewed.

REEXPRESSION OF VARIABLES

An alternative approach is to reexpress nonnormally distributed continuous variables so as to produce distributions that more closely approximate normality. The reexpressed variables can then be analyzed using normal theory estimation techniques (e.g., GLS) without producing biased estimates of model fit or the standard errors of the relationships between the reexpressed variables.

Item Parcels. A commonly used simple method of reexpression is the construction of item parcels by summing or taking the mean of several items that purportedly measure the same construct (e.g., Marsh, Antill, & Cunningham, 1989). These parcels will typically exhibit distributions that more closely approach a normal distribution than the original items. Another perhaps less obvious advantage of item parcels is that fewer parameters will need to be estimated in the measurement model, implying that the estimates will be more stable in small samples.

Note, however, that the construction of item parcels is not without its potential drawbacks (Cattell & Burdsal, 1975). Of most importance, the construction of parcels may obscure the fact that more than one factor may underlie any given item parcel. This problem leads to considerable potential complication in the interpretation of relationships and structure in models using item parcels. Moreover, the use of too few measured variables (parcels) as indicators of a construct yields

less stringent tests of the proposed structure of confirmatory factor models. Identification problems are also more likely to occur if too few item parcels are used per factor (i.e., < 3). In such cases, if the correlation between factors is near 0, the model will not be identified.

Transformation of Nonnormal Variables. A transformation performs an operation on observed scores that preserves the order of the scores but alters the distance between adjacent scores. Linear transformations (e.g., standardization) have no effect on either the distributions of variables or the results of simple structural equation models that do *not* impose equality constraints (see Cudeck, 1989). Nonlinear transformations potentially alter the distribution of the measured variables as well as the relationships among measured variables, potentially eliminating some forms of curvilinear effects and interactions between variables. In the presentation below, we assume that all observed values of the variable being transformed are greater than 0, a condition that can be achieved by adding a constant to each observation.

Two classes of approaches to selecting an appropriate transformation are available. First, a power function of the variable may be identified that produces a new (transformed) variable that more closely approximates normality. Several sources (e.g., Daniel & Wood, 1980) offer rules of thumb for selecting power transformations. Given positively skewed distributions, taking logarithmic, square root, or reciprocal transformations (or, more generally, raising the scores on the measured variable to a power less than 1.0) will typically result in distributions that more closely approximate normality. Given negatively skewed distributions, raising raw scores to a power greater than 1.0 will often result in a more normally distributed transformed variable. Daniel and Wood (1980) present plots that are highly useful in selecting a potential transformation. Emerson and Stoto (1983) present a useful technique, the transformation plot for symmetry, in which simple functions of scores associated with specified percentile ranks are plotted. The slope of the resulting graph helps identify the optimal power transformation.

A second class of approaches is useful when scatterplots suggest a possible nonlinear relationship between pairs of variables. Box and Cox (1964) suggested framing this as a nonlinear regression problem: The slope (b_1) and intercept (b_0) of a linear regression equation, $y^{\lambda} = b_0 + b_1 x + e$, are estimated simultaneously with the optimal power transformation (λ) for the dependent variable. In practice, several regression equations representing values of λ over the range -2 to $+2$ (with the

logarithmic transformation representing a value of 0) may be computed, selecting the value of λ for which R^2 is maximized as the optimal power transformation. A more recent exploratory approach, the Alternating Conditional Expectation (ACE) algorithm (see de Veaux, 1990), goes one step further, finding the transformation of each variable that produces the maximum *possible* R^2 between y and x (or even a set of predictors). The ACE algorithm finds optimal transformations that maximize the linear relations between two variables, even when power transformations are unsuccessful.

The Box-Cox and ACE approaches have considerable power when applied to single regression equations; however, structural equation analysts must recall that they are seeking a single transformation that is applicable across a series of regression equations, some of which involve latent variables. Consequently, Box-Cox and ACE must be viewed as providing guidance, rather than a definitive solution in the search for a single transformation that will improve the linearity of the set of relations involving an initially problematic variable.

Several observations should be made about transformations. First, the univariate skewness and kurtosis of the transformed data should always be examined to assess the improvement, if any, in the distribution of the new variable. These indices are also useful in choosing between competing transformations. Note that for some distributions of observed variables, there will be no simple power transformation that will substantially reduce the skewness and kurtosis. Second, the Mardia measures of multivariate skewness and kurtosis for the original and transformed variables should be compared for the set of original and transformed variables. Recall that well-behaved univariate distributions are only a necessary and not a sufficient condition for multivariate normality. Third, although the second approach to transformation, increasing the linearity of relationships, does not directly address normality, linearizing transformations often have the additional benefit of improving the distribution and homoscedasticity of errors of measurement. Fourth, transformation of the data changes the original measure y to a new measure y^*. The new correlations or covariances are computed between the y^* transformed variables, not between the original variables. Reflecting this change, fit statistics, parameter estimates, and standard errors will be based on the y^* variables and may differ, perhaps substantially, from those based on the original variables. Fifth, the application of the ACE algorithm to any measured variable or of different power transformations to each measured variable can potentially result in considerable confusion in the interpretation of the transformed results. Even more severe interpretational problems result when different transformations are applied to the *same* measured variable across studies. This is particularly problematic in the use of the ACE algorithm because of its strong tendency to capitalize on chance relationships that cannot be expected to replicate across studies. In general, the loss of metric associated with the transformation is an issue to the extent that researchers wish to compare results across variables or across studies. In addition, the original metrics of the measured variables may represent important units in some areas of social science (i.e., income in dollars). However, in other areas of social science, measures are more often assessed in arbitrary metrics (e.g., seven-point Likert scales), so it is less crucial to preserve the scale of measurement.

Conclusion and Recommendations

The effect of nonnormality on structural equation modeling depends on both its extent and its source (poorly distributed continuous variables, coarsely categorized variables, or outliers). In general, the greater the extent of nonnormality, the greater the magnitude of the problem. Our presentation above has detailed the statistical effects of each of the sources of nonnormality on χ^2 goodness-of-fit statistics, parameter estimates, and standard errors. These problems also have important practical implications. Researchers obtaining inflated χ^2 goodness-of-fit statistics because of nonnormal data will be tempted to make inappropriate, nonreplicable modifications in theoretically adequate models to achieve traditional standards of fit (MacCallum, Roznowski, & Necowitz, 1992; Chapter 2, this volume). Underestimated standard errors will produce significant paths and correlations between factors, even though they do not exist in the population. Such "findings" can be expected to fail to be replicated, contributing to confusion in many research areas.

The choice among the remedial measures again depends on the extent and source of the nonnormality, as well as the sample size. Considering first the measures of goodness of fit and standard errors for continuous, nonnormally distributed variables, both the ADF estimator and the Satorra-Bentler SCALED χ^2 and robust standard errors have shown very good performance, regardless of the degree of nonnormality in large samples when the model has been correctly specified.

What is meant by "large samples" has varied across studies, but it is clearly of the range of 1000 to 5000 cases. For sample sizes in the range of approximately 200 to 500 cases, depending on the extent of nonnormality, the Satorra-Bentler statistics appear to have the best properties. For smaller sample sizes, we recommend normal theory ML or GLS estimates when the distributions are not substantially nonnormal, and the Satorra-Bentler statistics as the distributions begin to depart substantially from normality (e.g., skewness = 2; kurtosis = 7). Under these conditions, the use of a more stringent level of α for tests of parameters might also be considered. Particularly for smaller sample sizes, we also recommend inspection of the CFI or Bollen's (1989a) IFI, which have only a small downward bias (3% to 4%), even under severely nonnormal conditions. Note that these recommendations assume that the model has been correctly specified.

For small sample sizes in particular, the two methods of reexpression considered here may improve normal theory estimation techniques. The construction of item parcels usually produces composite variables that more closely approximate normality. The data reduction accomplished in the process also yields a more favorable parameter-to-subject ratio, which is likely to be particularly important in small samples (Bentler & Chou, 1988). Transformations can also often yield new variables that more closely approximate normal distributions. Identification of the optimal normalizing transformation is less certain in small than in large samples. The identification of an adequate transformation that is satisfactory for normal theory estimation can be achieved in some, but not all, data sets. Each of the reexpression methods has its own disadvantages: Item parcels may obscure multifactorial structures; the loss of the original metric from transformation may complicate the interpretation of the results. To date, little empirical work has been done specifically investigating the effect of reexpression techniques on the results of structural equation modeling analyses.

Finally, the CVM estimator appears to provide the most appropriate estimates of the model χ^2, parameter estimates, and standard errors with coarsely categorized, skewed data. The primary advantage of the CVM estimator over the competing normal theory and the Satorra-Bentler statistics occurs as the number of ordered categories decreases. With five or more categories, there is little or no benefit to using CVM; with two categories, there is a substantial advantage given poorly distributed observed variables. The CVM estimator appears to produce good results for ordered categories only with large samples (e.g., at least

500-1000 depending on the complexity of the model being estimated). In addition, the relationships provided by the CVM estimator are between latent, normally distributed variables rather than the original measured variables, potentially complicating interpretation of the results. Coarsely categorized variables that cannot be conceptualized as having an underlying normal distribution, or for which latent variables cannot be constructed that have joint normal distributions, are not appropriate candidates for the technique and are likely not appropriate candidates for structural equation modeling.

The remedies prescribed here address the majority of situations in which nonnormality arises in practice. Most of the remedies are easy to program and are increasingly available as options in the standard computer packages for structural equation modeling (see Chapter 8, this volume). These advances will make it easier for researchers to check the distributional assumptions underlying normal theory estimation and to select and implement alternative approaches when they are not met.

to the memory of Jeffrey S. Tanaka

STRUCTURAL EQUATION MODELING

Concepts, Issues, and Applications

Rick H. Hoyle
editor

SAGE Publications
International Educational and Professional Publisher
Thousand Oaks London New Delhi

For information address:

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications Ltd.
6 Bonhill Street
London EC2A 4PU
United Kingdom

SAGE Publications India Pvt. Ltd.
M-32 Market
Greater Kailash I
New Delhi 110 048 India

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Structural equation modeling: concepts, issues, and applications / edited by Rick H. Hoyle
p. cm.
Includes bibliographical references and index.
ISBN 0-8039-5317-8 (acid-free). — ISBN 0-8039-5318-6 (pbk.: acid-free)
1. Social sciences—Mathematical models. 2. Social sciences—Statistical methods. I. Hoyle, Rick H.
H61.25.S767 1995
300'.1'5118—dc20 94-47262

This book is printed on acid-free paper.

95 96 97 98 99 10 9 8 7 6 5 4 3 2 1

Sage Project Editor: Susan McElroy

Brief Contents