Fitting Genetic Models with LISREL: Hypothesis Testing


A brief introduction to the mathematical theory involved in model fitting is provided. The properties of maximum-likelihood estimates are described, and their advantages in fitting structural models are given. Identification of models is considered. Standard errors of parameter estimates are compared with the use of likelihood-ratio (L-R) statistics. For structural modeling, L-R tests are invariant to parameter transformation and give robust tests of significance. Some guidelines for fitting models to data collected from twins are given, with discussion of the relative merits of parsimony and data description.

KEY WORDS: structural modeling; LISREL; maximum likelihood.

INTRODUCTION

Model fitting is a method of translating hypotheses into mathematical expressions which can be compared with a set of data. In order to make this comparison meaningful, certain conditions for identification (described below) should be met. If the model fits well, then the data are consistent with the model and, therefore, provide support for the model. If the model fits badly, then the model is said to be rejected in the light of the data.
of the data. However, there are a number of different methods of assessing the fit of models and of testing specific hypotheses. The relative merits of these approaches are discussed in this paper.

IDENTIFICATION OF PARAMETERS

The free parameters of a model may be said to be either (i) underidentified, (ii) just identified, or (iii) overidentified. Underidentification necessarily occurs if there are more parameters in the model than distinct expected statistics. For example, a model with additive genetic, common environment, and specific environment parameters is underidentified if we have data from monozygotic (MZ) twins reared together. The expected covariance matrix contains three statistics: the covariance between twins, the variance of twin 1, and the variance of twin 2. However, only two of these statistics are distinct under the model, because both variances have the same expectation under the model. A model may be just identified if the number of distinct expected statistics is equal to the number of observed statistics. Identification of this type is comparatively rare in behavioral genetics, where models are usually overidentified, as the twin or adopted research designs collect data containing several "replicate" statistics, such as the variances of twins. The overidentified model is fitted by making the expected statistics as "close" as possible to all the observed statistics; maximum likelihood (see below) is one scale of "closeness" or "goodness of fit."

Although it is possible for a model to be just identified or overidentified if the number of different expected statistics is equal to or greater than the number of parameters, it is not necessarily the case. If two or more parameters are completely confounded, it is not possible to obtain an unambiguous estimate of these parameters. For example, if we attempt to partition variation into additive ($h$) and dominance ($d$) genetic, common ($c$) and specific environment ($e$), from covariance matrices of pairs of siblings, half-siblings reared apart, and cousins, the number of observed statistics is nine (three covariances and six variances). There are four different expectations under the model:

\[ \text{Var} = h^2 + d^2 + c^2 + e^2, \]
\[ \text{Cov(sibs)} = .5h^2 + .25d^2 + c^2, \]
\[ \text{Cov(half-sibs)} = .25h^2 \]
\[ \text{Cov(cousins)} = .125h^2 \]

Although there are four parameters, and four equations, the expected resemblance of cousins is simply a linear function of the resemblance of
half-sibs. Hence an infinite number of combinations of parameter values will give the same fit of the model to the data. In these simple cases, the expectations are simple linear combinations of the parameters. An unweighted least-squares solution to the problem is obtained by solving the system of equations

\[ Ax = b, \]

where \( x \) is the vector of parameters, \( b \) is the vector of observed statistics (averaged if there are replicates), and \( A \) is a matrix of coefficients such that element \( ij \) is the weight of the \( j \)th parameter in the \( i \)th expected equation. If the matrix \( A \) is nonsingular, the model is identified, and an unweighted least-squares solution is obtained from

\[ x = A^{-1}b. \]

Thus our example above may be represented as

\[
\begin{bmatrix}
1 & 1 & 1 & 1 \\
.5 & .25 & 1 & 0 \\
.5 & 0 & 0 & 0 \\
.25 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
h^2 \\
d^2 \\
e^2
\end{bmatrix}
= \begin{bmatrix}
b_1 \\
b_2 \\
b_3 \\
b_4
\end{bmatrix}.
\]

The matrix \( A \) is clearly singular in this case, since the fourth row is simply half of the third row. Data on cousins do not add to the information from the variance and the covariances of sibs and half-sibs under this model.

This simple logic does not readily extend to cases in which nonlinear functions of parameters form part of the expected correlations. Nevertheless, considerable insight to the identification status of models is often obtained by examining the model in this way. For example, the identification of both the linear terms \( x \) and \( xy \) within a model is sufficient to conclude that both \( x \) and \( y \) are identified for nonzero values of \( x \).

Jöreskog and Sörbom (1986) state that there are "no general and practically useful necessary and sufficient conditions for identification are available for the general LISREL model." There are general conditions for identification, but their practical significance is not always great. A necessary condition for identification is that the number of observed statistics (i.e., variances and covariances) is greater than or equal to the number of parameters in the model. If there is the same number of statistics as parameters, the model may be just identified, and no degree of freedom is available to test the fit of the model. If there are more observed statistics than parameters, the model may be overidentified, and some measure of fit of the model can be obtained. A sufficient condition for identification is that each equation of a model is distinct from every other
equation in the model and from all possible linear combinations of the other equations in the model. In practice, with good minimization software, a numerical check on identification is possible. If we simulate data from the model, using a parameter vector \( \theta_1 \), then regardless of our starting values, if we fit the model to the simulated data, the estimated parameter vector should equal \( \theta_1 \). Unfortunately, LISREL can be sensitive to starting values and fail to obtain the correct solution if the starting values are a long way from the solution, so this test is not 100% reliable with this software.

LISREL VI provides a check of parameter identification by examining the positive definiteness of the information matrix. A warning message is printed on the output if the information matrix is found to be nonpositive definite. While this check is usually reliable, it occasionally gives spurious messages that certain parameters may not be identified when they are estimated to be close to zero. Greater detail and references concerning issues of identification can be found in the LISREL manual.

**ASSESSMENT OF FIT**

**The Likelihood-Ratio Test**

LISREL VI provides several different methods of fitting models to data. These methods are described in the LISREL manual (Jöreskog and Sörbom, 1986) so they are not described in detail here. Of the various methods available, maximum likelihood (ML) is by far the most popular for genetic applications. ML has a number of desirable properties (see Edwards, 1972; Cavalli-Sforza and Bodmer, 1971), many of which were discovered and elucidated by Fisher (1920, 1922). First, ML estimates of parameters are asymptotically unbiased—as sample sizes tend to infinity, bias disappears. Second, ML estimates are asymptotically most efficient—having the minimum variance of all asymptotically unbiased estimates and hence making the best use of the available data. Third, ML estimates are invariant to transformation, so, for example, the ML estimate of a parameter \( p \) would be the square root of the estimate of \( p^2 \) for a given set of data. Fourth, ML estimates are sufficient statistics, since they define the likelihood function up to an arbitrary constant. Sufficiency ensures that the estimates extract all the relevant information in the data. Mathematically, it is often convenient to fit models using the natural logarithm of the likelihood function.

It is useful to consider data as a sample from a probability distribution. The equation used to define the height of this distribution at a particular place on the \( x \) axis is known as the probability density function (p.d.f.). The p.d.f. describes data points in terms of the probability dis-
tribution when the parameters of the distribution are held constant. The likelihood employs the same mathematical function, but the data points are held constant and the parameters of the distribution vary. In behavioral genetics, it is frequently assumed that data are sampled from a multivariate normal distribution. Let the row vector \( x_i \) be one set of observations from a multivariate normal distribution. This set may be one individual measured on several different variables or several relatives measured on one or many variables. If there are \( k \) variables in the set, the p.d.f. corresponding to the observed vector \( x_i \) is

\[
\frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} e^{-1/2(x_i - \mu_i)^\prime \Sigma^{-1}(x_i - \mu_i)},
\]

where \( \Sigma \) is the population covariance matrix and \( \mu_i \) is the population mean vector of the variables. The logarithm of the likelihood is therefore

\[-\frac{1}{2} \{ k \log(2\pi) + \log |\Sigma| + (x_i - \mu_i)^\prime \Sigma^{-1}(x_i - \mu_i) \},\]

where \( |\Sigma| \) and \( \Sigma^{-1} \) denote the determinant and inverse of the matrix \( \Sigma \), respectively.

If we take a number of independent samples, the log-likelihoods may be summed to give a total log-likelihood for the whole sample. In practice, the calculation of the likelihood for each vector of observations during model fitting would prove very expensive in computer time. Fortunately, it is possible to simplify the general case in which all variables are measured in every individual ("balanced pedigrees" with no missing data). It can be shown (see, e.g., Mardia et al., 1979, p. 97) that the sum of the log-likelihoods of the \( n \) pedigrees is

\[
L_A = -\frac{n}{2} \{ k \log(2\pi) + \log |\Sigma| + \text{tr}(\Sigma^{-1}S) + (\bar{x} - \mu)^\prime \Sigma^{-1}(\bar{x} - \mu) \}.
\]

The maximum-likelihood estimate of the mean vector, \( \bar{x} \), is equal to the population mean vector, \( \mu \), so at the maximum, the term \( (\bar{x} - \mu)^\prime \Sigma^{-1}(\bar{x} - \mu) \) is zero, giving

\[
L_A = -\frac{n}{2} \{ k \log(2\pi) + \log |\Sigma| + \text{tr}(\Sigma^{-1}S) \}.
\]

McArdle (1986) and Dolan et al. (1989) describe genetic applications of fitting models to augmented moment matrices in which the structure of the means is predicted by the covariance structure. Here we confine our treatment to model fitting to variance–covariance matrices. If the ob-
served and the population matrices are equal ($\Sigma = S$), then the likelihood is simply

$$L_B = -\frac{1}{2} n [k \log(2\pi) + \log |S| + k].$$

Twice the difference between $L_A$ and $L_B$ is

$$F = n \left\{ \log \left( \frac{|\Sigma|}{|S|} \right) + \text{tr}(S\Sigma^{-1}) - k \right\},$$

which, if $n$ is large, is approximately distributed as $\chi^2$ and is a convenient function for LISREL to minimize. When there are $m$ independent matrices, $S_j$, each conforming to the Wishart distribution (as will be the case if the original observations are sampled from a multivariate normal distribution), the approximate $\chi^2$ value is given by

$$F = \sum_{j=1}^{m} n_j \left\{ \log \left( \frac{|\Sigma_j|}{|S_j|} \right) + \text{tr}(S_j\Sigma_j^{-1}) - k_j \right\}$$

and the degrees of freedom are

$$df = \left[ \sum_{j=1}^{m} \frac{1}{2} k_j (k_j + 1) \right] - x,$$

where $k_j$ is the number of variables in the $j$th matrix $S_j$ and $x$ is the number of free parameters in the model.

If the number of free parameters in the model is reduced (for example, by fixing parameters to a value or constraining parameters to equal functions of other parameters), the difference between the expectations of the new model and the observed data is still distributed as $\chi^2$. This is a very useful result, since by subtraction it is obvious that the difference between the $\chi^2$ of the initial model and the $\chi^2$ of the submodel is itself distributed as $\chi^2$, with degrees of freedom $= (df_{\text{model 1}} - df_{\text{model 2}})$. Therefore, a statistical test of specific hypotheses can be made. This test is statistically correct when the data are multivariate normally distributed and the sample size is fairly large. The approach is particularly useful when there are several elements of the covariance matrix which are expected to be equal, for example, when we have a covariance of twin 1 on variable A with twin 2 on variable B, and a covariance of twin 2 on variable A with twin 1 on variable B, or two parent–child covariances within a family. In this case, the method makes full use of the data, while taking into account the lack of statistical independence of the observations. As shown in the next section, the $\chi^2$ test is superior to the use of standard errors of parameters for testing hypotheses about covariance structure.
The Use of Standard Errors

Optionally, LISREL VI provides standard errors of estimates of parameters. These are intended to give the user an idea of the precision with which estimates are known and to provide a check on the identification of parameters. When the model is underidentified, estimates of standard errors of parameters cannot be computed since the Hessian matrix is singular. It is common practice to use the value of a parameter estimate divided by its standard error in order to assess whether the estimate differs significantly from zero. In this section, we demonstrate that while the significance of parameters is unaffected by transformations when the likelihood-ratio statistic is used as a criterion, the same is not true of the (mis)use of standard errors.

When a statistic, \( x \), has a distribution which tends to normality in large samples, it can be shown (e.g., Fisher, 1922) that the variance is the inverse of the second derivative of the likelihood function with respect to that parameter. So if \( x \) is evaluated at \( \hat{x} \),

\[
\frac{1}{\sigma_x^2} = -\frac{\delta^2 L}{\delta x^2},
\]

where \( L \) is the sum of the log-likelihoods of the observations. Let

\[
c_x = \frac{\hat{x}}{\sigma_x} = \hat{x} \sqrt{\frac{\delta^2 L}{\delta x^2}}
\]

be a constant calculated from \( \hat{x} \), the maximum-likelihood estimate of \( x \).

Now consider \( g(x) \), a differentiable function of \( x \). In this case, since by the chain rule

\[
\frac{\delta L}{\delta x} = \frac{\delta L}{\delta g(x)} \cdot \frac{\delta g(x)}{\delta x},
\]

we have

\[
\frac{\delta L}{\delta g(x)} = \left[ \frac{\delta g(x)}{\delta x} \right]^{-1} \cdot \frac{\delta L}{\delta x},
\]

so the second derivative of the log-likelihood with respect to the function of \( x \) is

\[
\frac{\delta^2 L}{\delta g(x)^2} = \left[ \frac{\delta g(x)}{\delta x} \right]^{-1} \cdot \frac{\delta^2 L}{\delta x \cdot \delta g(x)} + \frac{\delta [\delta g(x)/\delta x]}{\delta g(x)} \cdot \frac{\delta L}{\delta x}.
\]
Now $\delta L / \delta x = 0$ when $L$ is maximized with respect to $x$, so the second term on the right-hand side disappears. We note that because

$$
\frac{\delta^2 L}{\delta x^2} = \frac{\delta^2 L}{\delta x \cdot \delta g(x)} \cdot \frac{\delta g(x)}{\delta x},
$$

the second derivative of the log-likelihood with respect to $g(x)$ becomes

$$
\left( \frac{\delta g(x)}{\delta x} \right)^{-2} \cdot \frac{\delta^2 L}{\delta x^2}.
$$

Hence the constant in the case of $g(x)$ is

$$
c_{g(x)} = \hat{g}(x) \sqrt{\frac{\delta^2 L}{\delta g(x)^2}} = \hat{g}(x) \cdot \left( \frac{\delta g(x)}{\delta x} \right)^{-1} \cdot \sqrt{\frac{\delta^2 L}{\delta x^2}}.
$$

For example, if $g(x) = x^2$, then $c_{g(x)}$ will be half of $c_x$. Thus the use of $c$ is not recommended for the assessment of statistical significance of parameter estimates.

For most applications, the likelihood-ratio test using the $\chi^2$ statistic is to be preferred. In cases involving two or more parameters, estimates of which may be correlated, the relationship between the standard errors of functions of parameters may be more complex. The interpretation of standard errors becomes more difficult, especially when the first derivatives of the function with respect to the parameters are zero (see Kendall and Stuart, 1977, pp. 246–247). The likelihood-ratio test is still valid in these more complex situations and, despite being more difficult to apply, is the preferred method of assessing the significance of parameter estimates.

**Alternative Measures of Fit**

While there are clear advantages of fitting models by ML and using $\chi^2$ goodness-of-fit statistics, the method is not without problems. As pointed out in the LISREL manual, the $\chi^2$ statistic is valid only if the variables have a multivariate normal distribution, the model is fitted to covariance, not correlation, matrices, and the sample size is fairly large. In behavioral genetics, the second and third assumptions are often met, but the first is not. Unfortunately, the $\chi^2$ statistic is sensitive to both sample size and departures from normality. In large samples quite small departures from normality or effects accounting for a very small proportion of the total variance can lead to a significant $\chi^2$. As a result, several alternative indices of fit have been proposed.\(^3\)

\(^3\) Since the preparation of our manuscript, an excellent review of several goodness-of-fit indices has been provided by Marsh et al. (1988).
LISREL provides several means of assessing fit—the residual matrices (O–E), the goodness-of-fit index (GFI), and the root mean square residual (RMR). The GFI is independent of sample size and is given for each matrix in the model. A perfect fit is given by a GFI of unity, while a poor fit is suggested by values less than, say, .8. Boomsma et al. (1989) employ this statistic to assess the fit of models for sports participation and heart rate measured in twins and their parents. A similar summary of the data is provided by the Tucker and Lewis (1973) reliability coefficient. A modification of this coefficient was used by Martin et al. (1979) to address the case of multiple groups. If we divide the $\chi^2$ statistic by $n$, the degrees of freedom, the result is an $F$ statistic with $n$ and $\infty$ degrees of freedom. The expected value of the $F$ statistic is 1 if the model fits, so $F - 1$ gives a measure of goodness of fit of the model. If we calculate $F_0$, the statistic for a null model (e.g., no covariation between family members), and $F_1$, the statistic for a model with additional parameters, then the coefficient

$$
\rho = \frac{F_0 - F_1}{F_0 - 1}
$$

gives the improvement in fit as a fraction of the total improvement possible. The RMR is more difficult to use as a yardstick, since it depends on the sizes of the observed variances and covariances, but is basically an average of the elements of the residual matrix. A further possibility is to examine the normalized residuals, noting poor fit to statistics for which the normalized residual is greater than two. This approach is used by Martin and Boomsma (1989) to compare the fit of a Markovian and a factor model of the personality and willingness to drive when drunk.

Clearly, the more parameters used in a model, the closer the agreement between observed and expected matrices is likely to be. It would be of little value to fit fully saturated models which simply describe the data in a different way. There is a need for a statistic which reflects not only the goodness of fit of a model, but also its simplicity. A popular and easy-to-calculate measure in Akaike’s information criterion (AIC; Akaike, 1970, 1987). For applications fitting to variance–covariance matrices,

$$
\text{AIC} = \chi^2 - 2(\text{df}).
$$

A fully saturated model generally has $\chi^2 = 0$ and $\text{df} = 0$ so the AIC is zero. An improvement on this situation is reflected in a negative value of AIC, and the best model is often the model with the lowest AIC. Examples of the use of the criterion are given for factor analysis by Akaike
(1987) and for a multivariate model of genetic and cultural transmission by Phillips et al. (1987).

**Interpretation of Goodness of Fit**

We emphasize that a good fit of a model to some data does not prove that the model is correct. On the contrary, the model could be completely and utterly false yet give a good account of the observed data. A poor fit of a model to a set of data is in some ways more informative, since the model is shown to be incorrect (although note the sensitivity of the $\chi^2$ statistic in large samples). In order to test competing models, it is necessary to collect data which afford some resolution of the alternative hypotheses. A simple example would be if data were collected from MZ twins but not dizygotic (DZ) twins. A model of purely environmental effects would not be rejected with these data, as any twin correlation could be due to shared environmental effects. The addition of DZ twin data may reject this hypothesis if the DZ twin correlation is significantly lower than that of the MZ twin. The same is true of the new set of data in the sense that while a good fit may be obtained with a genetic model, it is possible that a different genetic or environmental architecture is responsible for variation. For example, it is not possible to resolve genetic dominance and common environment effects with data collected from twins reared together (although a model involving one source of variation may have more support from the data than the other).

The principle of parsimony, or "Ockham's razor," embodies the idea articulated by Leibnitz that the most productive theories are "simplest in hypotheses and richest in phenomena." The principle usually leads to simpler theories which are more easily tested and have a greater scope and, thus, a greater chance of being accurate. When assessing the fit of models and submodels, parsimony dictates that parameters with estimates not significantly different from zero should be dropped from the model. Vlietinck et al. (1989) use this method in the analysis of birth-weight data and fit a parsimonious model without covariation between maternal and gestational age and with no common environmental effects. As a result, the estimates of the other parameters of the model have smaller standard errors than if the nonsignificant parameters had been retained. This method runs the risk of making a type II error by assuming sources to be absent when they are in fact present, thus leading to biases in the estimates of parameters. Despite this risk, it is still the method of choice, since if the assumption that the cause is absent is correct, more rapid progress may be made when trying to partition variation further. The method is self-correcting, so that if false assumptions have been made,
then it is likely that the further collection of data will reveal a weakness of the model and force rejection of the initial assumptions. Everyone likes to be right, but to attempt perfection by fitting saturated models without seeking simpler explanations amounts to noninductive empiricism. The history of science suggests that the most significant gains are not won without risk and that to be wrong is sometimes to be most productive.

CLASSICAL TWIN STUDIES

In the case of univariate studies of twins reared together, a number of possible models may, in principle, be tested with the data. One of the most complex models would be to allow for additive genetic (VA), common environment (VCE), specific environmental (VSE), sibling phenotypic interaction effects, scalar sex differences in all parameters, and non-scalar sex differences in genetic variability. It is hoped that the world is not so complex as this, and one may start by testing the simplest of hypotheses—a model of random environmental variation that is the same in both sexes. Typically, this model is rejected, and additional parameters are included. If there are no a priori grounds for expecting genetic or common environmental effects, it is logical to fit a model of VA, VSE, and VCE and to examine the fit. This model is often the first to be fitted to any set of twin data, as it provides a simple yet powerful structure for explanation. If the model fits, then it may be possible to drop one of the two new parameters and leave a good account of the data, particularly if an estimate is small.

When the three-parameter model does not fit, several additional levels of complexity may be added. Often the next step is to test for sex differences in parameter estimates. Perhaps the simplest of these elaborations is to assume merely a difference in total variance between the sexes, which can be modeled by allowing the parameters for one sex to be a constant multiple of parameters for the other sex. Such a model is not very easy to specify using LISREL but may be achieved using the $B$ matrix (see Heath et al., 1989; Neale and Martin, 1989). Different VA, CE, and SE parameters for the two sexes may be specified if total variance differences do not provide an adequate fit. If this model fails also, we may follow with a model of nonscalar sex limitation (i.e., different genes or different common environments for the two sexes) and then by a model incorporating reciprocal interaction.

The order of fitting models is not inflexible. For example, sibling interaction effects may be expected to account for a large proportion of variance of some variables and, hence, may be included at an earlier stage than sex-limitation effects. In order to get a good “feel” for the data, the
investigator may wish to fit a complete set of models to the data and then to draw as parsimonious a conclusion as possible. However, some caution is required with judging significance levels of $\chi^2$ associated with the large number of tests, as the possibility for type I error increases.

**FUTURE DEVELOPMENTS**

At the time of writing, LISREL VII is still in preparation, but it is worthwhile to note some of the additional features which will become available in the near-future. Classical theory of covariance structures assumes multivariate normality of the variates. Browne (1984) has described an extension of this classical theory to any multivariate distribution of continuous variables. The idea is to provide estimates of the variances and covariances of the elements of the variance–covariance data matrices. These statistics (based on fourth-order central moments) may be calculated with PRELIS and are used for a weight matrix in a generalized least-squares loss function. The method (known as asymptotically distribution-free best generalized least squares; AGLS) is implemented in a LISREL-like program called LISCOMP (Muthén, 1988) and in Bentler’s (1984) structural modeling program EQS. However, EQS has no provision for multiple groups, an essential feature for the applications described in this issue. Under very general assumptions about distributions, AGLS allows a test of the overall goodness of fit of models and a comparison of submodels using the $\chi^2$ statistic. However, the number of elements of the weight matrix rapidly escalates as the order of the covariance matrix increases. Nevertheless, this new approach is likely to become popular in many areas of the social sciences where continuous, normally distributed variables are the exception rather than the rule.

**REFERENCES**


Hypothesis Testing with LISREL.


