History of LISREL

Most first, and even second, courses in applied statistics seldom go much further than ordinary least squares analysis of data from controlled experiments, group comparisons, or simple prediction studies. Collectively, these procedures make up regression analysis, and the linear mathematical functions on which they depend are referred to as regression models. This basic method of data analysis is quite suitable for curve-fitting problems in physical science, where an empirical relationship between an observed dependent variable and a manipulated independent variable must be estimated. It also serves well the purposes of biological investigation in which organisms are assigned randomly to treatment conditions and differences in the average responses among the treatment groups are estimated.

An essential feature of these applications is that only the dependent variable or the observed response is assumed to be subject to measurement error or other uncontrolled variation. That is, there is only one random variable in the picture. The independent variable or treatment level is assumed to be fixed by the experimenter at known predetermined values. The only exception to this formulation is the empirical prediction problem. For that purpose, the investigator observes certain values of one or more predictor variables and wishes to estimate the mean and variance of the distribution of a criterion variable among respondents with given values of the predictors. Because the prediction is conditional on these known values, they may be considered fixed quantities in the regression model. An example is predicting the height that a child will attain at maturity from his or her current height and the known heights of the parents. Even though all of the heights are measured subject to error, only the child's height at maturity is considered a random variable.

Where ordinary regression methods no longer suffice, and indeed give misleading results, is in purely observational studies in which all variables are subject to measurement error or uncontrolled variation and the purpose of the inquiry is to estimate relationships that account for variation among the variables in question. This is the essential problem of data analysis in those fields where experimentation is impossible or impractical and mere empirical prediction is not the objective of the study. It is typical of almost all research in fields such as sociology, economics, ecology, and even areas of physical science such as geology and meteorology. In these fields, the essential problem of data analysis is the estimation of structural relationships between quantitative observed variables. When the mathematical model that represents these relationships is linear, we speak of a linear structural relationship. The various aspects of formulating, fitting, and testing such relationships we refer to as structural equation modeling.

Although structural equation modeling has become a prominent form of data analysis only in the last twenty years (thanks in part to the availability of the LISREL program), the concept was first introduced nearly eighty years ago by the population biologist, Sewell Wright, at the University of Chicago. He showed that linear relationships among observed variables could be represented in the form of so-called path diagrams and associated path coefficients. By tracing causal and associational paths on the diagram according to simple rules, he was able to write down immediately the linear structural relationship between the variables. Wright applied this technique initially to calculate the correlation expected between observed characteristics of related persons on the supposition of Mendelian inheritance. Later, he applied it to more general types of relationships among persons.

The modern form of linear structural analysis includes an algebraic formulation of the model in addition to the path diagram representation. The two forms are equivalent and the implementation of the analysis in the LISREL program permits the user to submit the model to the computer in either representation. The path analytic approach is excellent when the number of variables involved in the relationship is moderate, but the diagram becomes cumbersome when the number of variables is large. In that case, writing the relationships symbolically is more convenient. The SIMPLIS manual presents examples of both representations and makes clear the correspondence between the paths and the structural equations. Notice that in the above-mentioned fields in which experimentation is hardly ever possible, psychology and education do not appear. Controlled experiments with both animal and human subjects have been a mainstay of psychological research for more than a century, and in the 1920s experimental evaluations of instructional methods began to appear in education. As empirical research developed in these fields, however, a new type of data analytic problem became apparent that was not encountered in other fields.

In psychology, the difficulty was, and still is, that for the most part there are no well-defined dependent variables. The variables of interest differ widely from one area of psychological research to another and often go in and out of favor within areas over relatively short periods of time. Psychology has been variously described as the science of behavior or the science of human information processing. But the varieties of behavior and information handling are so multifarious that no progress in research can be made until investigators identify the variables to be studied and the method of observing them. Where headway has been made in defining a coherent domain of observation, it has been through the mediation of a construct-some putative latent variable that is modified by stimuli from various sources and in turn controls or influences various observable aspects of behavior. The archetypal example of such a latent variable is the construct of general intelligence introduced by Charles Spearman to account for the observed positive correlations between successful performance on many types of problem-solving tasks.

Investigation of mathematical and statistical methods required in validating constructs and measuring their influence led to the development of the data analytic procedure called factor analysis. Its modern form is due largely to the work of Truman Kelly and L.L.Thurstone who transformed Spearman's one-factor analysis into a fully general multiple-factor analysis. More recently, Karl Jöreskog added confirmatory factor analysis to the earlier exploratory form of analysis. The two forms serve different purposes. Exploratory factor analysis is an authentic discovery procedure: it enables one to see relationships among variables that are not at all obvious in the original data or even in the correlations among variables. Confirmatory factor analysis enables one to test whether relationships expected on theoretical grounds actually appear in the data. Derrick Lawley and Karl Jöreskog provided a statistical procedure, based on maximum likelihood estimation, for fitting factor models to data and testing the number of factors that can be detected and reliably estimated in the data.

Similar problems of defining variables appear in educational research, even in experimental studies of alternative methods of instruction. The goals of education are broad and the outcomes of instruction are correspondingly many: an innovation in instructional practice may lead to a gain in some measured outcomes and a loss in others. The investigator can measure a great many such outcomes, but unless all are favorable or all unfavorable the results become too complex to discuss or provide any guide to educational policy. Again, factor analysis is a great assistance in identifying the main dimensions of variation among outcomes and suggesting parsimonious constructs for their discussion.

In the LISREL model, the linear structural relationship and the factor structure are combined into one comprehensive model applicable to observational studies in many fields. The model allows multiple latent constructs indicated by observable explanatory (or exogenous) variables, recursive and nonrecursive relationships between constructs, and multiple latent constructs indicated by observable responses (or endogenous) variables. The connections between the latent constructs compose the structural equation model; the relationships between the latent constructs and their observable indicators or outcomes compose the factor models. All parts of the comprehensive model may be represented in the path diagram and all factor loadings and structural relationships appear as coefficients of the path.

Nested within the general model are simpler models that the user of the LISREL program may choose as special cases. If some of the variables involved in the structural relationships are observed directly, rather than indicated, part or all of the factor model may be excluded. Conversely, if there are no structural relationships, the model may reduce to a confirmatory factor analysis applicable to the data in question. Finally, if the data arise from a simple prediction problem or controlled experiment in which the independent variable or treatment level is measured without error, the user may specialize to a simple regression model and obtain the standard results of ordinary least-squares analysis.

These specializations may be communicated to the LISREL computer program in three different ways. At the most intuitive, visual level, the user may construct the path diagram interactively on the screen and specify paths to be included or excluded. The corresponding verbal level is the SIMPLIS command language. It requires only that the username the variables and declare the relationships among them. The third and most detailed level is the LISREL command language. It is phrased in terms of matrices that appear in the matrix-algebraic representation of the model. Various parameters of the matrices may be fixed or set equal to other parameters, and linear and non-linear constraints may be imposed among them. The terms and syntax of the LISREL command language are explained and illustrated in the LISREL program manuals. Most but not all of these functions are included in the SIMPLIS language; certain advanced functions are only possible in native LISREL commands.

The essential statistical assumption of LISREL analysis is that random quantities within the model are distributed in a form belonging to the family of elliptical distributions, the most prominent member of which is the multivariate normal distribution. In applications where it is reasonable to assume multivariate normality, the maximum likelihood method of estimating unknowns in the model is justified and usually preferred. Where the requirements of maximum likelihood estimation are not met, as when the data are ordinal rather than measured, the various least squares estimation methods are available. It is important to understand, however, except in those cases where ordinary least squares analysis applies or the weight matrices of other least squares methods are known, that these are large-sample estimation procedures. This is not a serious limitation in observation studies, where samples are typically large. Small-sample theory applies properly only to controlled experiments and only when the model contains a single, univariate or multivariate normal error component.

The great merit of restricting the analytical methods to elliptically distributed variation is the fact that the sample mean and covariance matrix (or correlation matrix and standard deviations) are sufficient statistics of the analysis. This allows all the information in the data that bear on the choice and fitting of the model to be compressed into the relatively small number of summary statistics. The resulting data compression is a tremendous advantage in large-scale sample-survey studies, where the number of observations may run to the tens of thousands, whereas the number of sufficient statistics are of an order of magnitude determined by the number of variables.

The operation of reducing the raw data to their sufficient statistics (while cleaning and verifying the validity for the data) is performed by the PRELIS program which accompanies LISREL. PRELIS also computes summary statistics for qualitative data in the form of tetrachoric or polychoric correlation matrices. When there are several sample groups, and the LISREL model is defined and compared across the groups, PRELIS prepares the sufficient statistics for each sample in turn.

In many social and psychological or educational research studies where a single sample is involved, the variables are usually measured on a scale with an arbitrary origin. In that case, the overall means of the variables in the sample can be excluded from the analysis, and the fitting of the LISREL model can be regarded simply as an analysis of the covariance structure, in which case the expected covariance matrix implied by the model is fitted to the observed covariance matrix directly. Since the sample covariance matrix is a sufficient statistic under the distribution assumption, the result is equivalent to fitting the data. Again, the analysis is made more manageable because one can examine the residuals from the observed covariances, which are moderate in number, as opposed to analyzing residuals of the original observations in a large sample.

Many of these aspects of the LISREL analysis are brought out in the examples in the PRELIS and LISREL program manuals. In addition, the SIMPLIS manual contains exercises to help the student strengthen and expand his or her understanding of this powerful method of data analysis. Files containing the data of these examples are included with the program and can be analyzed in numerous different ways to explore and test alternative models.

Today, however, LISREL is no longer limited to SEM. The latest LISREL for Windows includes the following statistical applications.

- LISREL for structural equation modeling.
- PRELIS for data manipulations and basic statistical analyses.
- MULTILEV for hierarchical linear and non-linear modeling.
- SURVEYGLIM for generalized linear modeling.
- MAPGLIM for generalized linear modeling for multilevel data.

LISREL has a set of 12 accompanying PDF user guides that can be accessed via the Help menu of the application.