Please download the examples from our website at https://ssicentral.com/index.php/products/lisrel/lisrel-examples and unzip them into a convenient folder location. The examples shown in the manual uses ‘C:\LISREL Examples’ and you are more than welcome to use the same or a different location. Please note that the actual location may be different on your machine.
Table of Contents

1 Introduction ............................................................................................................................................. 4

2 Complex Sampling Concepts ............................................................................................................. 4
   2.1 Introduction ........................................................................................................................................... 4
   2.2 Indicator variables and t-estimators ................................................................................................. 5
   2.3 Additional weight adjustments ........................................................................................................... 8
   2.4 Linear regression ................................................................................................................................. 9
   2.4.1 Example 1 ...................................................................................................................................... 9
   2.4.2 Example 2 ...................................................................................................................................... 9
   2.5 Standard error estimation ................................................................................................................... 12
   2.6 Heteroscedastic model ....................................................................................................................... 14
   2.7 Covariance matrix of vector of totals ............................................................................................... 15
   2.7.1 Introduction .................................................................................................................................. 15
   2.7.2 Notation ....................................................................................................................................... 15
   2.7.3 Total covariances ......................................................................................................................... 16
   2.8 Approximate covariance matrix of estimators ............................................................................... 19
   2.9 References ......................................................................................................................................... 22

3 Structural equation models for continuous outcomes ........................................................................... 24
   3.1 Introduction ........................................................................................................................................ 24
   3.1.1 The structural equation model ..................................................................................................... 26
   3.2 Graphical User Interface ................................................................................................................... 28
   3.2.1 The new PTH window .................................................................................................................. 28
   3.2.2 The Setup menu ............................................................................................................................ 30
   3.2.3 The Title and Comments dialog box ............................................................................................ 31
   3.2.4 The Group Names dialog box ....................................................................................................... 31
   3.2.5 The Labels dialog box .................................................................................................................. 32
   3.2.6 The Data dialog box ..................................................................................................................... 33
   3.2.7 The graphic pane of the PTH window ......................................................................................... 35
   3.2.8 The Weight Cases and Survey Design dialog boxes .................................................................. 36
   3.3 Examples ........................................................................................................................................... 39
   3.3.1 A structural equation model for the 2001 Monitoring the Future data ........................................... 39
   3.3.2 Implementation of sampling weights in a linear growth curve model ......................................... 54
   3.3.3 Implementation of sampling weights in a linear growth curve model ......................................... 59
   3.3.4 Incorporating stratum and cluster variables in the model ............................................................ 71
   3.4 Evaluation ......................................................................................................................................... 75
   3.4.1 Simulation study based on a linear growth curve model ............................................................... 75
   3.4.2 Latent curve analysis with main and interaction effects ................................................................. 78
   3.4.3 Replicate weights ......................................................................................................................... 93
   3.5 Theory ............................................................................................................................................... 106
   3.5.1 Introduction .................................................................................................................................. 106
   3.5.2 Parameter estimation for continuous manifest variables ............................................................ 106
   3.6 References ......................................................................................................................................... 115

4 Structural equation models for a mixture of ordinal and continuous outcomes ...................... 117
   4.1 Introduction ...................................................................................................................................... 117
4.2 A re-parameterization of the general LISREL model.................................................................................................................................. 117
4.3 Mixture of ordinal and continuous variables.................................................................................................................................. 119
4.4 Parameter estimation.................................................................................................................................................................................. 121
4.5 Approximate covariance matrix of estimators.................................................................................................................................. 123
4.6 Adjustment to the chi-square goodness of fit statistics.................................................................................................................................. 124
4.7 Examples........................................................................................................................................................................................................ 125
  4.7.1 Confirmatory factor analysis model.......................................................................................................................................................... 125
  4.7.2 Confirmatory factor analysis model with latent variable relationship and latent variable means .................................................. 130
4.8 References and further reading........................................................................................................................................................................... 135
1 Introduction

There has been a growing interest in recent years in fitting models to data collected from longitudinal surveys that use complex sample designs. This interest reflects expansion in requirements by policy makers and researchers for in-depth studies of social processes over time. Traditionally, the analyses of complex survey samples have been carried out using specialized software packages. More recently, a number of statistical analyses packages, for example SAS and SPSS, have implemented procedures to handle complex survey data appropriately in the case of regression models with continuous and categorical outcome variables. In this guide we describe techniques currently implemented in LISREL for analyzing complex surveys data. Research on the longitudinal analysis of complex survey data with LISREL was supported by SBIR grant R43 AA014999-01 from NIAAA to Scientific Software International.

A common theme in substance use research is that data are usually obtained from a multi-stage or so-called complex sampling data. A complex sampling design typically entails stratification, often on the basis of geography; defining meaningful clusters of population elements (PSUs); and one or more stages of subsampling within each PSU. While a complex sample has the advantages of being more economical and practical, guarantees a better representative sample of the population, and does not require a complete sampling frame of the population elements, it is generally less efficient than simple random sampling.

2 Complex Sampling Concepts

2.1 Introduction

A common theme in alcohol abuse research is that data are usually obtained from a multi-stage or complex sample design. An example of a typical complex sampling scheme is:

- Stratify the geographical area under study according to census geography and census socio-economic variables.
- Form meaningful clusters of population elements, called primary sampling units (PSUs), for example schools, in each stratum.
- Draw a predetermined number of PSUs from each stratum, using probability sampling proportional to size.
- Do one or more stages of subsampling within each PSU.
- Draw a simple random sample of ultimate sampling units (USUs) at the last stage.

The main advantages of a complex sample (CS) in comparison with a simple random sample (SRS) are:

- CS does not require a complete sampling frame of the population elements.
- CS is more economical and practical.
- CS guarantees a representative sample of the population.
- CS makes a step-by-step design of the sample possible.

The main disadvantage of CS is that it is generally less efficient than SRS, i.e., it yields estimates of lower precision for a fixed sample size.

In the application of CS, the design effect (deff) and sampling weights play an important role. The design effect is defined as

\[
\text{deff} = \frac{\text{Variance of an estimate using complex sampling}}{\text{Variance of an estimate under SRS}}
\]

The design effect (deff) provides a rough and ready method of estimating the variance of survey statistics and of adjusting the output of standard statistical software packages for the complex sampling design. This aspect of deff derives from its assumed portability. See Kish (1965) for a discussion of design effects.

The deff is used not only to produce estimates of variance, but also to adjust the output of standard analyses. For example, the practitioner may utilize standard statistical software packages to conduct a regression analysis of a hypothesized linear relationship between survey variables, or to formulate a multi-way table and conduct a $\chi^2$ test of independence between survey variables. The output of standard statistical software packages gives wrong answers for such problems (because the underlying assumptions of the methods are not satisfied for complex survey designs). A first-order correction may be obtained by dividing the corresponding test statistic by the estimated deff. See Rao & Scott (1981), and Skinner, Holt, & Smith (1989).

We start with some known results for the sum of random variables and for multiple linear regression to illustrate adjustments that must be made to accommodate complex sampling properly. In Sections 2 and 3 we provide a brief summary of important concepts in complex sampling and in Sections 4 to 7 discuss how these concepts are currently applied to fit regression models to survey data. We will show that standard software packages for regression analysis allow for a weight variable, but do not yield the correct standard error estimates and measures of fit.

### 2.2 Indicator variables and t-estimators

Consider a finite population of identifiable units $U = \{u_1, u_2, \ldots, u_N\}$ where the size $N$ of the population is assumed known. The inclusion of a given element $u_i$ in a sample $s$ is a random event indicated by the random variable $I_i$ (sample membership indicator of element $i$), $i = 1, 2, \ldots, N$ defined as
\[ I_i = \begin{cases} 1 & \text{if } i \in s \\ 0 & \text{else.} \end{cases} \]

The probability that \( u_i \) will be included in the sample, denoted by \( \pi_i \), is

\[ \pi_i = P(u_i \in s) = P(I_i = 1). \]

The probability that both \( u_i \) and \( u_j \) will be included in the sample, denoted by \( \pi_{ij} \), is

\[ \pi_{ij} = P(u_i \in s \text{ and } u_j \in s) = P(I_i, I_j = 1). \]

By definition (see e.g. Traat, Meister, & Söstra, 2001). Therefore, \( E(I_i) = \pi_i \), \( Var(I_i) = \pi_i (1 - \pi_i) \) and \( \text{Cov}(I_i, I_j) = \pi_{ij} - \pi_i \pi_j \).

The selected sample \( s = \{u_{(1)}, u_{(2)}, \ldots, u_{(n)}\} \), is an unordered set of population units, where \( n \) denotes the sample size. A sampling weight \( w_i \) for the \( i \)-th USU is usually calculated as \( 1/\pi_i \), where \( \pi_i \) denotes the inclusion probability. If \( \pi_1 = \pi_2 = \ldots = \pi \), the sample is called a self-weighting sample. Sometimes \( w_i \) is called the base weight.

Let \( y_N \) and \( z_N \) be \( N \times 1 \) vectors of finite population values with typical elements \( y_j \) and \( z_j \), \( j = 1, 2, \ldots, N \) respectively. Denote the values drawn from a multi-stage sample of size \( n \) by \( y_s \) and \( z_s \), where

\[ y_s = (y_{(i)}, y_{(2)}, \ldots, y_{(j)}, \ldots, y_{(n)}), \]
\[ z_s = (z_{(i)}, z_{(2)}, \ldots, z_{(j)}, \ldots, z_{(n)}). \]

Here \( z_{(j)} \) denotes the \( j \)-th sample element, \( z_{(j)} \in z_s \).
Consider the population totals \( t_i = \sum_{i=1}^{N} y_i, \quad t_2 = \sum_{i=1}^{N} y_i^2, \quad t_3 = \sum_{i=1}^{N} z_i, \quad t_4 = \sum_{i=1}^{N} z_i^2, \) and \( t_5 = \sum_{i=1}^{N} z_i y_i. \) Each population total \( t_j \) can be estimated by the corresponding \( \pi \)-estimator \( \hat{t}_{j\pi} \) (Horvitz & Thompson, 1952). For example,

\[
\hat{t}_{1\pi} = \sum_{i=1}^{n} y(t) / \pi_i, \quad \hat{t}_{5\pi} = \sum_{i=1}^{N} z_i y_i / \pi_i.
\]

Each estimated total can be written as a linear function of the sample membership indicators \( I_i, i = 1, 2, ..., N. \) For example,

\[
\hat{t}_{1\pi} = \sum_{i=1}^{N} I_i y_i / \pi_i.
\]

This estimator (Horvitz-Thompson) is an unbiased estimator of \( t_i \) since

\[
E(\hat{t}_{1\pi}) = \sum_{i=1}^{N} E(I_i) y_i / \pi_i = \sum_{i=1}^{N} y_i = t_i.
\]

Use of \( \text{Cov}(I_i, I_j) = \pi_{ij} - \pi_i \pi_j \) gives

\[
\text{Var}(\hat{t}_{1\pi}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{1}{\pi_{ij}} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j.
\]

Hence,

\[
\text{Var}(\hat{t}_{1\pi}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{1}{\pi_{ij}} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j.
\]

This simple example shows that the expected value and variance of the sum are rather different under complex sampling than for the corresponding case of simple random sampling.

Standard methods are available for the point estimation of sample functions of the population totals, such as means, ratios, and differences of ratios. These methods are based (see e.g. Särndal, et. al., 1992) on the following result. Given that a population parameter \( \theta \) can be expressed as a function of
several population totals, \( \theta = f(t_1, t_2, \ldots, t_q) \), then an estimator \( \hat{\theta} \) of \( \theta \) is obtained from 
\[
\hat{\theta} = f\left(\hat{t}_{1\pi}, \hat{t}_{2\pi}, \ldots, \hat{t}_{q\pi}\right)
\]
where \( \hat{t}_{j\pi} \) is the corresponding \( \pi \)-estimator of \( t_j \). Additionally, consistent estimators of the sample variances of the estimators are available and have been implemented in various programs for the analysis of survey data (cf. Section 2.2). One method of estimating the variance of \( \hat{\theta} \) if \( \theta \) is a nonlinear function of the totals is by a first-order Taylor approximation of 
\[
f(h_{1\pi}, h_{2\pi}, \ldots, h_{q\pi})
\]
(see e.g. Wolter, 1985).

To estimate the variance of a survey estimator in the case of a single-stage sampling design, there are typically two alternatives: (1) the variance estimator based upon pps sampling with replacement, and (2) the Yates-Grundy estimator of variance (Yates & Grundy, 1953; Biyani, 1980) for pps without replacement sampling. Many survey practitioners will find the first estimator to be a satisfactory approximation to the variance given the actual survey design. For instances in which it is important to reflect the without replacement sampling design (e.g., important sampling fraction) and where it is feasible to calculate the joint inclusion probabilities (e.g., Durbin’s two-per-stratum design; see, Durbin, 1967, and Shapiro & Bateman, 1978), you would have the opportunity to specify the Yates-Grundy estimator. Applied to a multi-stage sampling design, these estimators usually provide a very good approximation to the total variance.

2.3 Additional weight adjustments

It was previously mentioned that the sampling weight \( w_i \) is usually calculated as \( w_i = 1/\pi_i \), the so-called base weight.

In a practical application each weight \( w_i \) usually undergoes additional adjustments,

- such as \( w_i = \frac{1}{\pi_i \hat{R}_i} \) (nonresponse adjusted weight), where \( \hat{R}_i \) is the response rate in a cell containing the \( i \)-th unit
- \( w_i = \frac{1}{\pi_i \hat{R}_i} F_i \) (post stratification weight), where \( F_i = \frac{T_k}{\hat{T}_k} \), and where \( T_k \) denotes the total number of units belonging to the \( k \)-th cell, \( k = 1, 2, \ldots, K \) of a contingency table formed from a set of categorical variables. For example, consider the variables age (5 categories), gender (2 categories), and ethnic group (3 categories). In this case \( K = 5 \times 2 \times 3 = 30 \). From the sample one can establish, for each of these variables, which category (e.g., male/female) is assigned to the \( i \)-th USU, and hence determine the cell number for that specific combination.

The estimated total \( \hat{T}_k \) is obtained using
\[ w_j = \frac{1}{\pi_i R_i}. \]

It is evident that non-response and post stratification adjusted weights will have an impact on the estimation of population totals and functions of population totals.

### 2.4 Linear regression

In this section we show the effect of sampling on the estimates and variability of regression coefficients. Again, the results are rather different from those that can be derived under the more familiar case of simple random sampling.

Suppose \( Y_N \) is an \( N \times p \) matrix defined as \( Y_N = (y_1, y_2, \ldots, y_N) \), where the elements of \( y_i \) are values of \( p \) variables of interest.

#### 2.4.1 Example 1

Let \( y_{ij} \) denote a typical element of \( y_i \), where \( y_{ij} \) is the number of occasions alcohol was consumed by student \( i (i = 1, 2, \ldots, N) \) in the prior 30 days, and where \( j = 1 \) denotes grade 8, \( j = 2 \) denotes grade 9, \( \ldots \), \( j = 5 \) denotes grade 12.

#### 2.4.2 Example 2

Let the subscript \( i \) denote the student \( i (i = 1, 2, \ldots, N) \). Suppose \( y_{i1} \) equals the number of times this student was under the influence of alcohol in the prior year; \( y_{i2} \) is a language score, and \( y_{i3} \) is a math score.

Example 1 above describes a longitudinal study, often referred to in the literature as a repeated measurements study, since measurements are made on the same individual on successive occasions. Note that, in general, measurement occasions are not necessarily equally spaced over time.

Example 2 describes a typical cross-sectional study. Note, however, that this study may have been carried out in 1998, and subsequently repeated in 2000 and 2002. It is evident that the finite populations \( U_{1998}, U_{2000}, \) and \( U_{2002} \) will overlap if, for example, 8th to 12th graders in the state of Texas are defined as the population elements. Hence, the samples \( s_{1998}, s_{2000}, \) and \( s_{2002} \) may also have overlapping units. A cross-sectional study, repeated over time, is often referred to as a panel study, but the statistical treatment usually treats the data as multiple-group data. In this examples the
year of study defines the group. A typical multiple group application is to test for differences in the means of latent variables under the assumption of factor invariance.

Consider the case $p = 1$ (univariate regression) so that $Y_N = y_N$, an $N$-dimensional vector. Suppose further that $X_N$ is an $N \times r$ matrix defined as $X'_N = (x_1, x_2, \ldots, x_N)$, where the elements of $x_i$ are values of $r$ auxiliary variables, for example $x_{i1} = \text{gender}$, $x_{i2} = \text{socio-economic status}$, and $x_{i3} = \text{age}$.

The finite population regression coefficient vector $\beta$ is a function of $y$ and is defined as

$$\beta = (X'_N X_N)^{-1} X'_N y_N. \quad (2.1)$$

Under a so-called design-based approach, $\beta$ is an obvious choice for the parameter of interest when regression is based on sample survey data. In the estimation of $\beta$ we assume an underlying homoscedastic model

$$E(y_N \mid X_N) = X_N \beta; \quad \text{Cov}(y_N \mid X_N) = \sigma^2 I_N. \quad (2.2)$$

Let $X_s$ denote an $n \times r$ matrix of rows of $X_N$ selected according to some sampling design $s$.

The ordinary least squares (OLS) estimator $\hat{\beta} = (X'_s X_s)^{-1} X'_s y_s$ of $\beta$ is not, in general, a design-consistent estimator of $\beta$.

An equivalent expression for (2.1) is

$$\beta = T^{-1} t \quad (2.3)$$

Where

$$T = \sum_{\alpha=1}^{N} x_{\alpha} x'_\alpha \quad \text{and} \quad t = \sum_{\alpha=1}^{N} x_{\alpha} y_\alpha.$$

Let $t_0$ and $t_{i0}$ denote typical elements of $T$ and $t$ respectively, then
\[ t_{ij} = \sum_{\alpha=1}^{N} x_{ia} x_{j\alpha} \quad \text{and} \quad t_{i0} = \sum_{\alpha=1}^{N} x_{ia} y_{\alpha}. \]

Each total (cf. Section 2.2) can be estimated by their unbiased \( \pi \)-estimators. For example,

\[ \hat{t}_{ij, \pi} = \sum_{\alpha=1}^{n} x_{(\alpha i)} x_{(j\alpha)} / \pi_{\alpha}, \]
\[ \hat{t}_{i0, \pi} = \sum_{\alpha=1}^{n} x_{(\alpha i)} y_{(j\alpha)} / \pi_{\alpha}, \quad i, j = 1, 2, \ldots, r. \]

In matrix notation,

\[ \hat{T} = \sum_{\alpha=1}^{n} \frac{x_{(\alpha)\prime} x_{(\alpha)}}{\pi_{\alpha}} = X\prime_{s} W_{s} X_{s}, \]
\[ (2.4) \]

And

\[ \hat{t} = \sum_{\alpha=1}^{n} \frac{x_{(\alpha)\prime} y_{(\alpha)}}{\pi_{\alpha}} = X\prime_{s} W_{s} y_{s}, \]
\[ (2.5) \]

where \( x_{(\alpha)} \) denotes the \( \alpha \)-th column of \( X_{s} \).

This yields the design-weighted estimator

\[ \hat{\beta}_{w} = \hat{T}^{-1} \hat{t} = (X\prime_{s} W_{s} X_{s})^{-1} X\prime_{s} W_{s} y_{s}, \]
\[ (2.6) \]

which is a design-consistent estimator of \( \beta \).

The weighting matrix is defined as \( W_{s} = \text{diag}(w_{1}, w_{2}, \ldots, w_{n}) \), where \( w_{i} = 1 / \pi_{i} \) denotes the sampling weight for the \( i \)-th USU. In the case of a self-weighting design, i.e., \( \pi_{i} = n / N, i = 1, 2, \ldots, N \), the estimators \( \hat{\beta} \) and \( \hat{\beta}_{w} \) are identical.
2.5 Standard error estimation

For most sample designs used in practice the sampling variance of $\hat{\beta}_w$ cannot be estimated using standard computer packages, and a variance estimating technique has to be used.

The basic methods available (see e.g. Rao, 1975) are:

(a) Linearization or Taylor expansion methods (Wolter, 1985, Binder, 1983):
Suppose that the parameter $\theta$ is a nonlinear function $f(t_1, t_2, \ldots, t_q)$ of population totals, then $\theta$ is consistently estimated by $\hat{\theta} = f\left(\hat{t}_{1,\pi}, \hat{t}_{2,\pi}, \ldots, \hat{t}_{q,\pi}\right)$. By Taylor linearization it follows that

$$f\left(\hat{t}_{1,\pi}, \hat{t}_{2,\pi}, \ldots, \hat{t}_{q,\pi}\right) \approx f\left(t_{1,\pi}, t_{2,\pi}, \ldots, t_{q,\pi}\right) + \sum_{j=1}^{q} a_j \left(t_{j,\pi} - t_j\right),$$

where

$$a_j = \frac{\partial f\left(\hat{t}_{1,\pi}, \hat{t}_{2,\pi}, \ldots, \hat{t}_{q,\pi}\right)}{\partial t_{j,\pi}} \bigg|_{t_{1,\pi} = t_1, \ldots, t_{q,\pi} = t_q}$$

From (1.3) it follows that $\beta$ is a nonlinear function of $r(r+1)/2+r$ population totals, with corresponding estimator $\hat{\beta}_w$ as defined in (2.6).

Using a first-order Taylor approximation, it can be shown that (cf. (2.3))

$$\hat{\beta}_w \approx \beta + T^{-1}\left(t - T\beta\right).$$

(2.7)

It follows that

$$\text{cov}(\hat{\beta}_w) \approx T^{-1}VT^{-1}.$$
An approximate expression for $T^{-1}VT^{-1}$ is $T^{-1}\hat{V}T^{-1}$. Typical elements of $V$ and $\hat{V}$ are given in, for example, Särndal et. al. (1992, page 194).

(b) Balanced repeated replication (McCarthy, 1969): Statistics based on half-samples, which are selected so as to ensure an orthogonal balanced set, are computed and the empirical covariances of these statistics are used as the appropriate estimator.

In longitudinal studies an increase in precision is obtained if allowance is made for the fact that units sampled over time are correlated. Replication methods provide a simple means for incorporating this correlation.

(c) Jackknife (Miller, 1974): The sample is first split into subsamples, each of which reflects the original complex design. Statistics based on the sample data without one of the subsamples are computed and the empirical covariances of these statistics serve as covariance estimators. A more detailed account is given in Wolter (1985).

(d) Bootstrap (Efron, 1981, 1982; Kovar, Rao & Wu, 1988): The sample data is used to construct an artificial population $U^*$ which is assumed to mimic the real, but unknown, population $U$. The original design is used to draw a series of $K$ samples (with replacement) from $U^*$. For each “bootstrap” sample, $i$, an estimate $\hat{\theta}_i^*$ of the population parameter $\theta$ is computed and subsequently $\hat{\theta}$ and $\text{var}(\hat{\theta})$ are estimated from $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_K^*$.

While the standard statistical package computer programs do not in general deal with the complex sample design situation, several special purpose programs for covariance estimation have been developed for use with complex sample designs. Lepkowski and Bowles (1996) give a list of eight software packages that are available for use by the general survey analyst. The eight catalogued in their paper are CENVAR, CLUSTERS, EpiInfo, PC CARP, STATA, SUDAAN, VPLX and WesVar. The first six programs use the Taylor series expansion for variance estimation and the remaining programs use replication methods.

Theoretical comparisons of the different methods of covariance estimation by Krewski & Rao (1981) and empirical comparisons by Kish & Frankel (1974) and by Richards & Freeman (1980) indicate their performance is very similar in many cases.
2.6 Heteroscedastic model

Some predictors may exhibit heteroscedasticity, in which changes in variance occur with changes in the values of the predictor. For example, the variance of income across individuals is systematically higher for higher-income individuals.

In the case of a heteroscedastic model

\[ E(y_N \mid X_N) = X_N \beta \quad \text{Cov}(y_N \mid X_N) = \sigma^2 \mathbf{V}, \]

(2.8)

the finite population parameter

\[ \beta^* = (X_N \mathbf{V}^{-1}X_N)^{-1} X_N \mathbf{V}^{-1}y_N \]

(2.9)

is a more suitable parameter for inference.

If \( \mathbf{V} \) is diagonal and the inclusion probabilities are proportional to the variances, then \( \hat{\beta}_w \) coincides with the weighted least squares estimator

\[ \hat{\beta}^* = (X_s \mathbf{V}_s^{-1}X_s)^{-1} X_s \mathbf{V}_s^{-1}y_s \]

(2.10)

where \( \mathbf{V}_s \) is the appropriate \( n \times n \) submatrix of \( \mathbf{V} \).

Standard programs compute the OLS estimator, \( \hat{\beta} \), and can often also compute the generalized OLS estimator, \( \hat{\beta}^* \), together with unbiased estimators of their model-variances \( \sigma^2(X'_sX_s)^{-1} \) under (2.2) and \( \sigma^2(X'_sV'_sV_s^{-1}X_s)^{-1} \) under (2.8) respectively. The design-weighted estimator, \( \hat{\beta}_w \), can also be obtained by the weighted regression options of standard statistical packages (e.g. LISREL or SPSS) by using the weights \( 1/\pi_i \). Alternatively, \( \hat{\beta}_w \) can be obtained by unweighted regression on the transformed variables \( y_i/\sqrt{\pi_i} \) and \( x_i/\sqrt{\pi_i} \). Nathan (1988), however, has pointed out that the reported variances and covariances will be incorrect. This implies that the standard significance tests (e.g. F-tests) will be invalid and can result in misleading conclusions.
The programs that use weighted regression, with weights \(1/\pi_i\), report the estimator of the variance-covariance matrix as \(\hat{\sigma}^2 \left(X_s'W_sX_s\right)^{-1}\). The model-variance of \(\hat{\beta}_w\), under the homoscedastic model (2.2), is

\[
\hat{\sigma}^2 \left(X_s'W_sX_s\right)^{-1}\left(X_s'W_sV_sW_sX_s\right)\left(X_s'W_sX_s\right)^{-1}
\]

which simplifies to \(\hat{\sigma}^2 \left(X_s'W_sX_s\right)^{-1}\) under the homoscedastic model \((V = I)\) only for self-weighting designs and under the heteroscedastic model (2.8) only if \(V\) is diagonal and the inclusion probabilities are proportional to the variances.

### 2.7 Covariance matrix of vector of totals

#### 2.7.1 Introduction

In this section formulae for the estimation of the covariance matrix of a vector of totals are given for single-stage, two-stage, and three-stage sampling designs.

For a multi-stage sampling design we assume the following general sampling methods at each stage:

- First stage: random sampling with replacement (WR), random sampling without replacement and equal probability of selection (WOR), and random sampling without replacement and unequal probabilities (UWOR).
- Second stage: if the first stage is not WR, then WR, WOR, or UWOR.
- Third stage: if second stage is not WR, then WR, WOR, or systematic.

From the above it follows that all specifications other than weights are ignored for subsequent stages if a multi-stage sample contains a WR, or an approximation to WR, stage.

Overall weights for each ultimate sampling unit can be obtained as a product of weights for corresponding units computed in each sampling stage.

#### 2.7.2 Notation

\(N\) : Total number of elements in the population
\( n \): Total number of elements in the sample
\( H \): Number of strata
\( n_h \): Sampled number of primary sampling units (PSU) per stratum
\( m_{hi} \): Number of elements in the \( i \)-th sampled PSU in stratum \( h \), \( i = 1, \ldots, n_h \)
\( w_{hi} \): Overall sampling weight for \( j \)-th element in the \( i \)-th sampled PSU in stratum \( h \)
\( y_{hi} \): Values of vector \( y \) for the \( j \)-th element in the \( i \)-th sampled PSU in stratum \( h \)
\( y_T \): Population total sum for vector of variables \( y \)

### 2.7.3 Total covariances

To simplify the expressions for the estimated covariance matrix of a vector of totals, let

\[
z_{hi} = w_{hi} y_{hi}
\]  \hspace{1cm} (2.11)

where the index \( h \) denotes a stratum within a given sampling stage, \( i \) denotes the \( i \)-th sampled unit within stratum \( h \) in the same sampling stage and \( j \) denotes all final stage units contained within \( hi \).

Let

\[
z_{hi} = \sum_{j=1}^{m_{hi}} z_{hij}
\]  \hspace{1cm} (2.12)

\[
\bar{z}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} z_{hi}
\]  \hspace{1cm} (2.13)

and

\[
S^2_h(y) = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (z_{hi} - \bar{z}_h)(z_{hi} - \bar{z}_h)
\]  \hspace{1cm} (2.14)

**Single stage sample**

The covariance of the total for vector \( y \) in a single-stage sample is estimated by:
\[ V \left( \hat{y}_T \right) = V_1 \left( \hat{y}_T \right) = \sum_{h=1}^{H} U_h \left( \hat{y}_T \right) \] 

(2.15)

where \( U_h \left( \hat{y}_T \right) \) is an estimated contribution from stratum \( h = 1, \ldots, H \) and depends on the sampling method used:

- For WR, \( U_h \left( \hat{y}_T \right) = n_h S_f^2 \left( y \right) \),
- For simple random sampling, \( U_h \left( \hat{y}_T \right) = (1 - f_h) n_h S_f^2 \left( y \right) \),
- and for sampling WOR and unequal probabilities,

\[
U_h \left( \hat{y}_T \right) = \sum_{i=1}^{n_h} \sum_{j > i}^{n_h} \left( \frac{\pi_{hi} \pi_{hj}}{\pi_{hj}} - 1 \right) \left( z_{hi} - \bar{z}_h \right) \left( z_{hj} - \bar{z}_h \right)
\]

In the variance estimator, \( \pi_{hi} \) and \( \pi_{hj} \) are the inclusion probability for units \( i \) and \( j \) in stratum \( h \), and \( \pi_{hij} \) is the joint inclusion probability for the same units (Yates & Grundy, 1953; Sen, 1953). In some situations it may yield a negative estimate and is treated as undefined.

Currently, for each stratum \( h \) containing a single element, the covariance contribution \( U_h \left( \hat{y}_T \right) \) is set to zero. An alternative procedure is to collapse strata. Presently, we leave it to the discretion of the user to collapse strata prior to any further statistical analysis.

**Two-stage sample**

When two-stage sampling is used and sampling WOR is applied in the first stage, the following estimate of the covariance of the total for vector \( y \) may be used:

\[ V \left( y_T \right) = V_2 \left( y_T \right) = V_1 \left( y_T \right) + \sum_{k=1}^{K_h} U_{hik} \left( y_T \right) \] 

(2.16)

- Here \( \pi_{hi} \) represents the first stage inclusion probability for the primary sampling unit \( i \) from stratum \( h \).
If simple random sampling is used, the inclusion probability is equal to the sampling rate $f_h$ for stratum $h$.

The number of second stage strata in the primary sampling unit $i$ within the first stage stratum $h$ is denoted by $K_{hi}$.

$U_{hk}(y_T)$ is the covariance contribution from the second stage stratum $k$ from the primary sampling unit $hi$. It depends on the sampling method used in the second stage (see formulae above).
Three-stage sample

For a three-stage sample where first stage sampling is done without replacement, and simple random sampling is applied in the second stage, the following estimate of the covariance of the total for vector $\mathbf{y}$ may be used:

$$
\mathbf{V}(\mathbf{y}_T) = \mathbf{V}_2(\mathbf{y}_T) + \sum_{h=1}^{H} \sum_{i=1}^{n_h} \pi_{hi} \sum_{j=1}^{K_h} f_{hik} \sum_{l=1}^{L_{hikl}} U_{hilk}(\mathbf{y}_T),
$$

(2.17)

where

- $f_{hik}$ represents the sampling rate for the secondary sampling units in the second-stage stratum $hik$,
- $L_{hik}$ indicates the number of third-stage strata in the secondary sampling unit $hikj$, and
- $U_{hilk}(\mathbf{y}_T)$ denotes the covariance contribution from the third-stage stratum $l$, which is contained in the secondary sampling unit $hikj$. Again, this depends on the third-stage sample method (see formulae above).

2.8 Approximate covariance matrix of estimators

In this section we provide a general procedure for the estimation of the approximate covariance matrix of estimators. The results derived are based on Binder (1983) and use a first-order Taylor linearization.

Assume that $L$ is the likelihood function or any other appropriate function of the vector $\mathbf{\gamma}$ of unknown parameters, and that an estimate $\hat{\mathbf{\gamma}}$ of $\mathbf{\gamma}$ is obtained by solving the set of simultaneous equations

$$
\frac{\partial \ln L}{\partial \mathbf{\gamma}} \bigg|_{\mathbf{\gamma}=\hat{\mathbf{\gamma}}} = 0
$$

(2.18)

In general, no closed-form solution to the set of equations (2.18) exists, and therefore parameter estimates are obtained iteratively using the Fisher scoring algorithm, for example,

$$
\hat{\mathbf{\gamma}}^{(t+1)} = \hat{\mathbf{\gamma}}^{(t)} + I^{-1}(\hat{\mathbf{\gamma}}^{(t)}) \mathbf{g}(\hat{\mathbf{\gamma}}^{(t)})
$$

(2.19)
where $\hat{\gamma}^{(t)}$ denotes the parameter values at iteration $t, t = 1, 2, \ldots, g(\cdot)$ denotes the gradient vector; and $I_n(\cdot)$ denotes the information matrix. In other words,

$$
g(\gamma) = \frac{\partial \ln L}{\partial \gamma} \quad (2.20)
$$

and

$$
I_n(\gamma) = -E \left[ \frac{\partial^2 \ln L}{\partial \gamma \partial \gamma'} \right] \quad (2.21)
$$

Denote the contribution to the gradient vector of each first-stage element for a given sampling stage by $g_{hij}$, where $h$ denotes stratum, and $i$ the $i$-th unit within this stratum. The index $j$ denotes a typical final stage element contained within the PSU $hi$, then

$$
[g(\gamma)]_r = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} [g_{hij}(\gamma)]_r \quad (2.22)
$$

From (2.18), (2.20) and (2.22) it follows that $\hat{\gamma}$ is the solution to the set of equations

$$
\hat{w}(\hat{\gamma}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} g_{hij}(\hat{\gamma}) = 0 \quad (2.23)
$$

Using a first-order Taylor expansion of $\hat{w}(\hat{\gamma})$ at $\hat{\gamma} = \gamma$, it follows that

$$
0 = \hat{w}(\hat{\gamma}) \approx \hat{w}(\gamma) + \frac{\partial \hat{w}(\gamma)}{\partial \gamma'} (\hat{\gamma} - \gamma) \quad (2.24)
$$

Taking variances on both sides, it further follows that

$$
Cov(\hat{w}(\hat{\gamma})) \approx \frac{\partial \hat{w}(\gamma)}{\partial \gamma'} Cov(\hat{\gamma}) \left( \frac{\partial \hat{w}(\gamma)}{\partial \gamma'} \right)' \quad (2.25)
$$

Thus, provided that (cf. (2.23))\( \frac{\partial^2 \ln L}{\partial \gamma \partial \gamma'} = \frac{\partial}{\partial \gamma} \left[ \frac{\partial g(\gamma)}{\partial \gamma'} \right] \) is a non-singular matrix,
\[
\text{Cov}(\hat{\gamma}) \approx \left[ \frac{\partial^2 \ln L}{\partial \gamma \partial \gamma'} \right]^{-1} \text{Cov}(\hat{w}(\hat{\gamma})) \left[ \frac{\partial^2 \ln L}{\partial \gamma \partial \gamma'} \right]^{-1},
\]

where \( E \left[ \frac{\partial^2 \ln L}{\partial \gamma \partial \gamma'} \right] = -I_n(\gamma) \).

Therefore, an approximate expression for the asymptotic covariance matrix of \( \hat{\gamma} \) is given by

\[
\text{Cov}(\hat{\gamma}) \approx I_n^{-1}(\gamma)G I_n^{-1}(\gamma)
\]

(2.26)

where \( G = \text{Cov}(\hat{w}(\hat{\gamma})) \).

Using results derived by Fuller (1975) (see also Section 2.7) it follows that, under single stage sampling with replacement (WR) or without replacement (WOR),

\[
G = \sum_{h=1}^{H} \frac{n_h (1 - f_h)}{n_h - 1} \sum_{i=1}^{n_h} (t_{hi} - \bar{t}_{hi})(t_{hi} - \bar{t}_{hi})',
\]

(2.27)

where:

- \( n_h = \sum_{j=1}^{n_h} m_{hij} \), with \( m_{hij} \) the number of cases with identical response patterns within stratum \( h \),
- \( f_h = \frac{n_h}{N_h} \), the sampling rate for stratum \( h \).
- \( t_{hi} = g_{hij}(\gamma) \) where \( g_{hij}(\gamma) \) is the \( h_{ij} \)-th contribution to the gradient vector \( g(\gamma) \) as defined by (2.19).
- \( t_{hi} = \frac{m_{hij}}{t_{hi}} \).
- \( t_{h} = \frac{1}{n_h} \sum_{i=1}^{n_h} t_{hi} \).
- \( t_{h} = \frac{1}{n_h} \sum_{i=1}^{n_h} t_{hi} \).

Currently, we assume a zero contribution to \( G \) for strata that contain a single PSU (cluster). Alternatively, the collapsing of strata or PSUs is left to the user’s discretion (see Section 2.7.3). Additionally, if there is no variable to define clusters, the observations within each stratum are treated as being the primary sampling units.
2.9 References


3 Structural equation models for continuous outcomes

3.1 Introduction

The single most important feature of the LISREL program is its facility to deal with a wide variety of models for the analysis of latent variables (LVs). In the social sciences, and increasingly in biomedical and public health research, LV models have become an indispensable statistical tool.

Because the whole framework of the LISREL model is based on relationships among LVs, it is worthwhile to briefly illustrate the concept of a latent variable. Latent variables are ubiquitous in some research domains, while in other contexts they are seldom used. In alcohol abuse studies, for example, they are a major focus of attention. It is the complexity of attitudes and traits underlying the alcoholism syndrome that is of greatest concern, rather than any specific behavior. As an example, questionnaire items are frequently collected that deal with the functioning of the subjects in a particular domain. Subsets of these items are often correlated. This implies that the subset reflects a common theme. For example, consider the following items from a hypothetical survey, used in a hypothetical model as shown in Figure 3.1:

- Q1: How many alcoholic drinks do you generally consume on any occasion?
- Q2: How many days in a typical week do you consume alcohol?
- Q3: Do you frequently attend parties where alcohol is available?
- Q4: Do you have alcoholic beverages with meals?

Figure 3.1: Path Diagram for Hypothetical SEM

In this example, a possible LV would be Tendency to Use Alcohol. This is a LV because Tendency to Use is a kind of unmeasurable propensity that is more than the combination of these items. The higher the individual Tendency LV score is, the more likely that the person will endorse questionnaire items regarding use and abuse of alcoholic beverages.

A LV is a statistical device used to summarize the information in a collection of correlated response variables. A LV describes the information of a set of items and reduces them to a single new measure. It is often assumed that the latent variable is superordinate to items on which it is based.
There are basically three major reasons for the utility of LV models. First, this kind of model can summarize information contained in many response variables by a few LVs. Consequently, the approach is parsimonious. Second, when properly specified a LV model can minimize the biasing effects of errors of measurement in estimating treatment effects. This means that the approach is often more accurate than is a traditional version of the same analysis. Third, LV models investigate effects between primary conceptual variables, rather than between any particular set of ordinary response variables. This means that a LV model is often viewed as more appropriate theoretically than is a simpler analysis with response variables only. A partial list of the sort of models that are subsumed under the framework of LISREL's general LV structure includes factor analysis, simultaneous equation models, standard growth curve processes, errors-in-variables models, virtually all forms of classical regression, univariate linear models and multivariate linear models, including the corresponding hypothesis tests on means and variances of classical experimental design. Literally hundreds of published articles appear each year that feature LV models, and an active program of statistical investigations on properties and extensions of LV models is carried out.

The hypothetical path diagram in Figure 3.2 shows seven $x$ variables as indicators of three latent $\xi$ variables. Note that $x_3$ is an indicator for both $\xi_1$ and $\xi_2$. There are two latent $\eta$ variables, each with two $y$ indicators. The model involves errors in equations (the $\zeta$ s), and errors in variables (the $\epsilon$ s and $\delta$ s).

**Figure 3.2: Path Diagram for Hypothetical SEM**

The LISREL model for single samples (Jöreskog & Sörbom, 1996) is defined by two components, namely the structural equation model and the measurement model(s).
3.1.1 The structural equation model

\[ \eta = \alpha + B\eta + \Gamma \xi + \zeta \]  

(3.1)

where \( \eta \) is a \( m \times 1 \) vector of endogenous latent variables and where it is assumed that the \( n \times 1 \) vector \( \xi \) of exogenous latent variables has mean \( \kappa \) and covariance matrix \( \Phi \), and that the \( m \times 1 \) vector \( \zeta \) of error terms has zero mean and covariance matrix \( \Psi \), and \( \text{cov}(\xi, \zeta) = 0 \). If \( |I - B| \neq 0 \), and setting \( A = (I - B)^{-1} \), it follows that

\[ \mu_\eta = A(\alpha + \Gamma \kappa) \]  

(3.2)

and

\[ \text{Cov}(\eta) = A(\Gamma \Phi \Gamma' + \Psi)A'. \]  

(3.3)

**Measurement models**

The measurement models for the \( p \) endogenous observed variables, represented by the vector \( y \), and the \( q \) exogenous observed variables, contained in the vector \( x \), relate the observed (manifest) variables to the underlying factors (latent variables) and may be expressed as

\[ y = \tau_y + \Lambda_y \eta + \varepsilon, \quad E(\varepsilon) = 0, \quad \text{Cov}(\varepsilon) = \Theta_\varepsilon \]

\[ x = \tau_x + \Lambda_x \xi + \delta, \quad E(\delta) = 0, \quad \text{Cov}(\delta) = \Theta_\delta \]

(3.4)

respectively.

The mean vectors of the observed variables are

\[ \mu_y = \tau_y + \Lambda_y A(\alpha + \Gamma \kappa), \quad \mu_x = \tau_x + \Lambda_x \kappa \]

(3.5)
In general, in a single population, $\tau_y$, $\tau_x$, $a$, and $k$ will not be identified without the imposition of further conditions. It further follows that

$$\Sigma_y = \Lambda_y \left[ A(\Gamma \Phi \Gamma^\prime + \Psi)A^\prime \right] \Lambda_y^\prime + \Theta_x$$

(3.6)

$$\Sigma_x = \Lambda_x \Phi \Lambda_x^\prime + \Theta_\delta$$

(3.7)

and

$$\Sigma_{yx} = \Lambda_y \Phi \Lambda_x^\prime.$$  

(3.8)

From (3.5) to (3.8), it follows that the covariance structure for the observed variables of the general LISREL model may be expressed as:

$$\Sigma = \text{Cov} \left[ \begin{bmatrix} y \\ x \end{bmatrix} \right] = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix}$$

(3.9)

From (3.5), the mean structure of the observed variables of the general LISREL model follows as:

$$\mu = E \left[ \begin{bmatrix} y \\ x \end{bmatrix} \right] = \begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix}.$$  

(3.10)

LISREL fits the mean-and-covariance structure defined in (3.9) and (3.10) to the data on the observed variables of the LISREL model. In this regard, LISREL can handle simple random sample data as well as complex survey data.

Special cases of the general LISREL model are obtained by fixing and constraining the parameters which are the elements in the 13 parameter matrices $\left( a, \kappa, \tau_x, \tau_y, \Lambda_y, \Lambda_x, B, \Gamma, \Phi, \Psi, \Theta_\epsilon, \Theta_\delta \right)$.

A large number of submodels is obtained by setting certain parameter matrices equal to the identity matrix or to zero. A few examples are:

- The measurement model for $x$, $x = \Lambda_x \xi + \delta$.  

27
A structural equation model where \( y \) and \( x \) are observed without error (\( A_y = I, A_x = I, \Theta_\zeta = 0, \Theta_\phi = 0 \))

\[
y = By + \Gamma x + \zeta
\]

Kaplan (2000) pointed out that this model was a major innovation in econometric modeling. In the special case where \( B = 0 \), one obtains the multivariate multiple regression model

\[
y = \Gamma x + \zeta
\]

The general form of the LISREL model, due to its flexible specification in terms of fixed and free parameters and simple equality constraints, has proven to be so rich that it can handle a large variety of problems. Using the inequality constraints feature in LISREL, users constantly discover new models, such as nonlinear growth curves (see du Toit & Cudeck, 2001) and vector time series models with ARMA residuals (du Toit & Browne, 2001) that can be handled within the LISREL framework.

There are many articles on structural equation modeling. Hayduk (1996), for example, gives a long list of substantive areas where structural equation models are being used: addictions, criminology, education, family studies, health, marketing, psychology, and sociology to mention just a few. A very large number of technical and substantive articles using structural equation models have appeared in dozens of journals.

### 3.2 Graphical User Interface

#### 3.2.1 The new PTH window

The path diagram component of the graphical user interface (GUI) of the LISREL module consists of the options and dialog boxes of the **Setup** menu on the PTH window of LISREL. This GUI component allows you to interactively generate the syntax file by means of a path diagram, which is a graphical representation of a structural equation model. The **Setup** menu on the PTH window of LISREL is reviewed in the next section while the four dialog boxes are reviewed separately in the subsequent sections. Thereafter, the use of the graphic pane of the PTH window is outlined.

The **Setup** menu on the PTH window provides access to a sequence of four dialog boxes that can be used to create a SIMPLIS or LISREL syntax file interactively by using a path diagram. A new PTH window is opened as follows. Open LISREL and select the **New** option on the **File** menu to create the following window.
Click on the **New** option on the **File** menu to load the **New** dialog box and select the **Path Diagram** option from the **New** dialog box as shown below.

Click on the **OK** button to load the **Save As** dialog box and then enter, for example, the name `demo.pth` in the **File name** field to produce the following dialog box.
3.2.2 The Setup menu

Next, click on the Save button to open the PTH window for demo.pth; and then click on the Setup menu to obtain the following window.

Typically, clicking on the Title and Comments option of the Setup menu will load the Title and Comments dialog box (see Section 3.2.2). However, you can click directly on the Groups, Variables or Data option to go to the Groups (see Section 3.2.3), the Variables (see Section 3.2.4), or the Data dialog box (see Section 3.2.5). Once you have completed the four sequential dialog boxes and drawn the path diagram, the SIMPLIS syntax file or the LISREL syntax file is generated by clicking on the Build SIMPLIS Syntax or the Build LISREL Syntax option respectively.
3.2.3 The Title and Comments dialog box

The Title and Comments dialog box allows you to specify a title and additional comments for the analysis. It is accessed by selecting the Title and Comments option on the Setup menu. This selection loads the following Title and Comments dialog box.

![Title and Comments Dialog Box]

Note that the Title and Comments dialog box corresponds with the Title command as shown above.

Once you are done with the Title and Comments dialog box, click on the Next button to go to the Group Names dialog box.

3.2.4 The Group Names dialog box

The Group Names dialog box is usually accessed by clicking on the Next button of the Title and Comments dialog box. It is required for multiple group analysis and allows you to specify different group names.

Note that the Group Names dialog box corresponds with the Group command as indicated on the image below. For single group analysis, you can skip this dialog box by simply clicking on the Next button.
Once the **Group Names** dialog box has been completed click the **Next** button to go to the **Labels** dialog box.

### 3.2.5 The Labels dialog box

The **Labels** dialog box allows you to specify the observed variables and latent variables of the model interactively. Access to this dialog box is obtained by clicking on the **Next** button of the **Group Names** dialog box. This selection loads the **Labels** dialog box as shown below.

Note that the **Labels** dialog box corresponds with the **Observed variables** and **Latent variables** commands as shown above. Note also that the **Add/Read Variables** dialog box, which is loaded by clicking on the **Add/Read Variables** button, corresponds with the **System file from file** and the **Raw data file from file** commands.

If a LISREL system file (DSF) or a Lisrel System file (LSF) is to be used, you can browse for the corresponding DSF or LSF by first selecting the **LISREL System File** option or the **Lisrel System File** option from the drop-down list box respectively and then clicking on the **Browse** button. Otherwise, you can add a list of variables by activating the **Add list of variables** radio button. When you are done with the **Add/Read Variables** dialog box, click the **OK** button to return to the **Labels** dialog box.

If the model includes any latent (unobservable) variables, you must specify labels for them by clicking on the **Add Latent Variables** button to load the **Add Variables** dialog box. Click the **OK** button after the label has been entered to return to the **Labels** dialog box. Once the labels for all the latent variables of the model have been specified, you click on the **Next** button to return to the **Data** dialog box.
3.2.6 The Data dialog box

Specify the data to be analyzed by using the Data dialog box. It is usually accessed by clicking on the Next button of the Labels dialog box. This action loads the following Data dialog box.
Note that the Data dialog box corresponds with the System file from file, Covariance matrix from file, Sample Size and Asymptotic covariance matrix from file commands as shown in the image above. If a DSF or a LSF is selected in the Labels dialog box and the covariance matrix is the matrix to be analyzed, the Data dialog box is redundant. In other words, in this case, you can click on the OK button to return to the PTH window without completing the Data dialog box.

In the case of a single-group analysis, the Groups drop-down list box is not accessible. In the case of a multiple group analysis, the Groups drop-down list box displays the labels of the different groups as specified in the Group Names dialog box. In this case, specify the data for each group by selecting the group name from the list box. If the latent variable means are to be compared across groups, click the Estimate latent means check box.

Select the desired data type from the Statistics from drop-down list box if the covariance matrix is not desired. Select the appropriate data file type from the File type drop-down list box if a DSF is not preferred and then use the Browse button to browse for the corresponding file. You can open the data file to be analyzed by clicking on the Edit button or specify the data by clicking on the New button. If the asymptotic covariance matrix or asymptotic variances of the sample moments is to be used in the analysis, you must check the Include weight matrix check box, select the type of weight matrix from the drop-down list box and browse for the appropriate file. If a DSF or a LSF is not to be processed, enter the...
number of observations in the **Number of observations** field. Select the desired moment matrix from the **Matrix to be analyzed** drop-down list box if a correlation matrix rather than a covariance matrix is to be analyzed.

After all four dialog boxes are completed, click on the **OK** button to return to the PTH window.

### 3.2.7 The graphic pane of the PTH window

Once you are done with the four dialog boxes of the **Setup** menu of the PTH window, the graphic pane of the PTH window is used to create a path diagram of the model to be fitted to the data. An example of the graphic pane of the PTH window is shown below.

You may use the following sequential steps to create a path diagram of the structural equation model to be fitted to the data in the graphic pane of the window.

- Use the **Select** tool to click, drag and drop the observed variables, one at a time, from the **Observed** list box into the graphic pane of the window.
- If the model includes latent variables, use the **Select** tool to click, drag and drop the latent variables, one at a time, from the **Latent** list box into the graphic pane of the window.
- Use the **One-way path** tool to specify the regression relationships between the observed and latent variables of the model.
- If applicable, use the **Two-way path** tool to specify the covariance (correlation) relationships between the latent variables and the error variables of the model.
- If certain parameters of the model are fixed to specific values, certain parameters are set equal to each other, a path needs to be removed from the model or certain graphic properties
are desired, the right-click menu of a path is used. This menu is activated by first selecting the path by using the **Select** tool and then by right-clicking to view the following menu.

<table>
<thead>
<tr>
<th>Fix</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔ Free</td>
</tr>
<tr>
<td>Set Value ...</td>
</tr>
<tr>
<td>Set Equal to ...</td>
</tr>
<tr>
<td>Cancel Setting Equal</td>
</tr>
<tr>
<td>Delete</td>
</tr>
<tr>
<td>Characteristics</td>
</tr>
<tr>
<td>Options...</td>
</tr>
<tr>
<td>Make the Line Straight</td>
</tr>
</tbody>
</table>

The options on the menu above are used as follows.
- The **Fix** option is used to fix a parameter that was set free by default.
- The **Free** option is used to free a parameter that was fixed by default.
- The **Set Value...** option is used to specify the value for a fixed parameter.
- The **Set Equal to...** option is used to set a parameter to be equal to another parameter(s).
- The **Cancel Setting Equal** option is used to release an equality constraint that was specified.
- The **Delete** option is used to specify the deletion of a path or a selected object.
- The **Characteristics** option is used to obtain information about the parameter.
- The **Options** option is used to modify the graphic properties of the path.
- The **Make Line Straight** option is used to automatically straighten a one-way path.

Once the path diagram has been drawn, click on the **Build SIMPLIS Syntax** option or **Build LISREL Syntax** option on the **Setup** menu to generate the SIMPLIS syntax file or the LISREL syntax file respectively.

### 3.2.8 The Weight Cases and Survey Design dialog boxes

The **Weight Cases** and **Survey Design** dialog boxes can be accessed if a LSF file is the active window. This is accomplished by selecting the **Data** menu from the main menu bar.
The Weight Cases option is used to calculate weighted sample statistics, for example means, covariances, and asymptotic covariances. It is assumed that these weights are normalized in the sense that the sum of sample weights equals the sample size.

The Survey Design dialog box shown below is used to define the stratification and cluster variables and to select a design weight. This information is stored within the LSF file and is retrieved whenever a SEM, based on an continuous outcome variable, is fitted to the data contained in the LSF file.
Presently the Survey Design feature is only available for structural equation models with continuous variables.
3.3 Examples

3.3.1 A structural equation model for the 2001 Monitoring the Future data

3.3.1.1 The data

The Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan has undertaken annual surveys designed to explore changes in important values, behaviors, and lifestyle orientations of contemporary American youth. The aims of these surveys are to provide a systematic, accurate description of the youth population of interest in a given year, and to explain relationships and trends observed over time. The Monitoring the Future surveys began in 1975. In the current example, data for 1608 respondents from the 2001 survey are used, and the focus is on relationships between the alcohol and marijuana use of respondents and traffic violations and/or accidents they were involved in.

Data for the first 10 participants on most of the variables used in this section are shown below in the form of a LSF named select.LSF which can be found in the TUTORIAL folder.

![Select LSF Table](image)

The following variables included in the LSF were selected from the survey data:
- school: This variable is used to indicate group membership of respondents within the 48 schools included in the survey.
- region: The 48 schools were drawn from 4 regions, and this variable indicates the region a school was drawn from.
- alclifs: The numerical response to the question "On how many occasions have you had alcoholic beverages to drink in your lifetime?"
- alc12mos: The numerical response to the question "On how many occasions have you had alcoholic beverages to drink in the past 12 months?"
- alc30ds: The numerical response to the question "On how many occasions have you had alcoholic beverages to drink in the past 30 days?"
o \textit{xmjlifs}: The numerical response to the question "On how many occasions have you used marijuana in your lifetime?"

o \textit{xmj12mos}: The numerical response to the question "On how many occasions have you used marijuana in the past 12 months?"

o \textit{xmj30ds}: The numerical response to the question "On how many occasions have you used marijuana in the past 30 days?"

o \textit{tick12mo}: The numerical response to the question "Within the last 12 months, how many times have you received a ticket (or been stopped and warned) for moving violations?"

o \textit{acci12mo}: The numerical response to the question "Within the last 12 months, how many times you were involved in an accident while driving?"

o \textit{newwgt}: The design weight of a student, computed as the inverse of the selection probability estimate of the region from which the student was selected. This selection probability estimate is merely the ratio of the sample frequency and the approximate population size of the region from which the student was selected.

3.3.1.2 The model

The five indicators or observed variables \textit{alc1ifs}, \textit{alc30ds}, \textit{xmjlifs}, \textit{xmj12mos}, and \textit{xmj30ds} are modeled to measure the alcohol and marijuana usage. Alcohol and marijuana usage, represented by the latent variables ALCUSAGE and MRJUSAGE in our proposed model, are modeled as causes of the number of moving violations and accidents, as represented by the \textit{Eta} variables ACCIDENT and TICKETS respectively. These two variables, in turn are measured without error by the two Y variables \textit{acci12mo} and \textit{tick12mo}. A path diagram of the model we intend fitting to the data is shown below.

![Path diagram of the model](image-url)
Mathematical Model

Measurement model

The measurement model for the latent variables ALCUSAGE, MRJUSAGE, ACCIDENT and TICKETS may be expressed as

\[
\begin{bmatrix}
    y \\
    x
\end{bmatrix} =
\begin{bmatrix}
    A_y & 0 \\
    0 & A_x
\end{bmatrix}
\begin{bmatrix}
    \eta \\
    \xi
\end{bmatrix} +
\begin{bmatrix}
    \varepsilon \\
    \delta
\end{bmatrix}
\]

where \( y = [ \text{tck21mo accl2mo} ] \), \( x = [ \text{alclifs alc30ds xmljifs xmlj12mos xmlj30ds} ] \), \( \eta = [ \text{TICKETS ACCIDENT} ] \), \( \xi = [ \text{ALCUSAGE MRJUSAGE} ] \), \( \delta = [ \delta_1 \delta_2 \delta_3 \delta_4 \delta_5 \delta_6 ] \), \( \varepsilon = [ \varepsilon_1 \varepsilon_2 ] \),

\[
A_y =
\begin{bmatrix}
    1 & 0 \\
    0 & 1
\end{bmatrix}
\]

and

\[
A_x =
\begin{bmatrix}
    \lambda_1 & 0 \\
    \lambda_2 & 0 \\
    0 & \lambda_3 \\
    0 & \lambda_4 \\
    0 & \lambda_5
\end{bmatrix}
\]

where \( \delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6, \varepsilon_1 \) and \( \varepsilon_2 \) denote measurement errors, and where \( \lambda_1, \lambda_2, \lambda_3, \lambda_4 \) and \( \lambda_5 \) denote unknown factor loadings.

Structural equation model

The structural equation model for the latent variables ALCUSAGE, MRJUSAGE, ACCIDENT and TICKETS is given by

\[
\eta = B\eta + \Gamma\xi + \zeta
\]

where \( \xi = [ \zeta_1 \zeta_2 ] \).
\[ B = \begin{bmatrix} 0 & \beta \\ 0 & 0 \end{bmatrix} \]

and

\[ \Gamma = \begin{bmatrix} \gamma_1 & \gamma_2 \\ \gamma_3 & \gamma_4 \end{bmatrix} \]

where \( \beta, \gamma_1, \gamma_2, \gamma_3 \) and \( \gamma_4 \) denote unknown regression weights, and \( \zeta_1 \) and \( \zeta_2 \) denote error terms.

The survey design variables school and region will be used as stratification and cluster variable respectively, while the design weight as represented by the variable newwgt will also be included in the specification of the analysis, as illustrated next.

### 3.3.1.3 Preparing the data and setting up the analysis

The model is fitted to the data in select.LSF by using the path diagram component (PTH window) of the LISREL GUI (See Section 3.2). After drawing the proposed model as a path diagram, SIMPLIS syntax is created and submitted. However, before we can fit the model, we need to specify the details of the complex survey design for the data in select.LSF.

The first step is to open the LSF, which is accomplished as follows:

- Use the File, Open option to activate the display of an Open dialog box.
- Set the Files of type drop-down list box to Lisrel Data (*.LSF) and browse for the file select.LSF in the TUTORIAL folder.
- Select the file and click the Open button to open the LSF in a LSF window.

### 3.3.1.3.1 Preparing the data

Click on the Survey Design option on the Data menu to load the Survey Design dialog box. Select the variable region from the Variables in data: list box and click on the Add button of the Stratification variable section. Next, select the variable school and add this variable to the Cluster variable section in a similar fashion. Finally, add the weight variable by selecting the variable newwgt from the Variables in data: list box and add this variable in the Design weight section. The completed Survey Design dialog box is shown below. Click on the OK button to return to the LSF window, and click the Save option on the File menu.
We now turn to creating a path diagram for the model to be fitted to these data. To open a new PTH window, select the **New** option on the **File** menu to load the **New** dialog box. Select the **Path Diagram** option from the list box on the **New** dialog box and provide a name for the path diagram, for example select.pth, in the **File name** string field of the **Save As** dialog box. Click on the **Save** button to open an empty PTH window.

Select the **Title and Comments** option on the **Setup** menu to load the **Title and Comments** dialog box. Enter the title A model for traffic tickets and accidents in the **Title** string field, and click on the **Next** button to load the **Group Names** dialog box.
Click on the **Next** button to load the **Labels** dialog box. Click on the **Add/Read Variables** button to load the **Add/Read Variables** dialog box, and select the **Lisrel System File** option in the **Read from file:** drop-down list box. Click on the **Browse** button to load the **Browse** dialog box and select the file **select.LSF** in the **TUTORIAL** folder. Click on the **OK** button to return to the **Labels** dialog box.

Click on the **Add Latent Variables** button to load the **Add Variables** dialog box. Enter the label **ALCUSAGE** in the string field and click **OK**. Enter the labels **MRJUSAGE**, **ACCIDENT**, and **TICKETS** in the same way.
The completed **Labels** dialog box is shown below. Click on the **OK** button to return to the PTH window for **select.pth**.

![Add Variables dialog box](image)

3.3.1.3.2 **Setting up the analysis**

At this point, an empty PTH window is displayed, with variable names listed to the left. Check the **Y** check boxes of `acci12mo` and `tick12mo` respectively. Check the **Eta** check boxes of `ACCIDENT` and `TICKETS` respectively to obtain the window shown below.
Next, click, drag and drop the labels of the Y variables one at a time into the PTH window. Position these variables to the right of the PTH window. Click, drag and drop the labels of the latent variables ACCIDENT and TICKETS one at a time into the PTH window to obtain the window shown below. Note that labels of variables dragged to the PTH window are shown against a colored background.
We now add the rest of the observed variables (alclifs, alc30ds, xmljifs, xmj12mos, and xmj30ds) one at a time into the PTH window, positioning them to the left of the PTH window. The last variables to be added are the latent variables ALCUSAGE and MRJUSAGE.

The next step is to add the paths between the variables dragged in the PTH window. Select the arrow icon on the Drawing toolbar, and click and drag indicator paths from the latent variable ALCUSAGE to alclifs and alc30ds respectively. To do so, start by clicking inside the ellipse representing ALCUSAGE and do not release the mouse button before the cursor is inside the rectangle representing alclifs or alc30ds. Click and drag similar indicator paths from the latent variable MRJUSAGE to xmljifs, xmj12mos and xmj30ds respectively.

Structural paths from the latent variable ALCUSAGE to both ACCIDENT and TICKETS, and from the latent variable MRJUSAGE to ACCIDENT and TICKETS are added in the same way. Also add indicator paths from the latent variable ACCIDENT to acci12mo, from the latent variable TICKETS to tick12mo, and from TICKETS to ACCIDENT. The model should now look like the image below.
The two indicator paths ACCIDENT to acc12mo, and TICKETS to tick12mo have to be fixed to a value of 1.0. To do so, deselect the arrow icon on the Drawing toolbar by clicking on the selection icon to its left. Next, right click on the path between ACCIDENT to acc12mo and select the Set Value option from the pop-up menu that appears. Set the value to 1.0 and click OK to return to the PTH window.

Right click on this path again, and select the Fix option from the pop-up menu. Note that the color representing the path has changed in the PTH window. Set the path between TICKETS to tick12mo to 1.0 in the same way.
Finally, set the error variances of acci12mo and tick12mo to zero by right-clicking the error arrows and selecting the **Fix** option from the pop-up menu.

The last paths to be added to the path diagram are the covariance between the measurement errors of xmj12mos and xmj30ds. Select the double arrow icon on the **Drawing** toolbar, and click and drag a path between the error arrows of xmj12mos and xmj30ds. Be sure to position the cursor over each arrow before activating and releasing the mouse button.

The path diagram should look like the following image.
Select the **Build SIMPLIS Syntax** option on the **Setup** menu. The generated syntax is automatically displayed in a SPJ window, as shown below.

![SPJ window](image1)

Click on the **Run LISREL** icon on the main toolbar to produce the following PTH window.

![PTH window](image2)
3.3.1.4 Discussion of results

Portions of the output file select.out are shown below.

From the results, it is evident that the five factor loadings are statistically significant if a 1% level of significance is used. In addition, the error covariance for xmj12mos and xmj30ds is also significant at a 1% level of significance. In other words, the results do not indicate any misspecifications in the measurement model of the latent variables ALCUSAGE and MRJUSAGE.

Since $\hat{\beta} = 0.38$ ($t = 11.03$), it follows that a student's number of accidents exerts a significant positive influence ($p < 0.01$) on his/her number of traffic tickets. Thus an increase in the number of accidents corresponds to an increase in the number of traffic tickets. Similarly, it follows that the alcohol usage of a student is a significant antecedent ($p < 0.01$) of both the number of accidents and the number of traffic tickets of the student. On the other hand, the marijuana usage is not a statistical significant antecedent of both the student’s number of accidents and traffic tickets.
The value for the number of accidents follows as $1 - 0.44 = 0.56$. In other words, the alcohol usage and marijuana usage of a student explains approximately 56% of the variation in the student’s number of accidents. Similarly, it follows that they explain approximately 27% of the variation in the number of traffic accidents of the student.
From the results above, it is evident that the $\chi^2$ test statistic value for the null hypothesis of a perfect fit is significant if a 1% level of significance is used. There is sufficient evidence that the theoretical model does not fit the data perfectly. However, the RMSEA point estimate of 0.0658 indicates that the model does provide a reasonable approximation to the data (Browne & Cudeck, 1993).
3.3.2 Implementation of sampling weights in a linear growth curve model

3.3.2.1 The data

A linear growth curve model with two dummy coded covariates (Lang1 and Lang2) is fitted to a simulated dataset contained in the LSF `surveysem.LSF`, contained in the `MISSINGEX` folder. It is assumed that the data are stratified according to 48 counties.

Within each county three schools are selected as primary sampling units (PSUs). In school number 1 four students are selected; in schools 2 and 3, three students are selected from each school. Students were selected on the basis of their initial achievement in an aptitude test (Score1) and measurements were repeated over six time intervals for five students from each school and over four time intervals for the remaining five.

The table below (Weight3) shows the weight calculations based on standardized initial scores.

<table>
<thead>
<tr>
<th>Interval</th>
<th>Lower</th>
<th>Upper</th>
<th>% Expected</th>
<th>% Selected</th>
<th>Weight3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-Inf</td>
<td>-1.00</td>
<td>15.87</td>
<td>10.00</td>
<td>1.587</td>
</tr>
<tr>
<td>2</td>
<td>-1.00</td>
<td>-0.70</td>
<td>8.33</td>
<td>10.00</td>
<td>0.833</td>
</tr>
<tr>
<td>3</td>
<td>-0.70</td>
<td>-0.20</td>
<td>17.88</td>
<td>10.00</td>
<td>1.788</td>
</tr>
<tr>
<td>4</td>
<td>-0.20</td>
<td>0.00</td>
<td>7.93</td>
<td>10.00</td>
<td>0.793</td>
</tr>
<tr>
<td>5</td>
<td>0.00</td>
<td>0.30</td>
<td>11.79</td>
<td>10.00</td>
<td>1.179</td>
</tr>
<tr>
<td>6</td>
<td>0.30</td>
<td>1.00</td>
<td>22.34</td>
<td>10.00</td>
<td>2.234</td>
</tr>
<tr>
<td>7</td>
<td>1.00</td>
<td>1.30</td>
<td>6.19</td>
<td>10.00</td>
<td>0.619</td>
</tr>
<tr>
<td>8</td>
<td>1.30</td>
<td>1.80</td>
<td>6.09</td>
<td>10.00</td>
<td>0.609</td>
</tr>
<tr>
<td>9</td>
<td>1.80</td>
<td>2.30</td>
<td>2.52</td>
<td>10.00</td>
<td>0.252</td>
</tr>
<tr>
<td>10</td>
<td>2.30</td>
<td>Inf</td>
<td>1.07</td>
<td>10.00</td>
<td>0.107</td>
</tr>
</tbody>
</table>

Ten students were selected from each school as follows:

- Four from racial group 1 with Weight2 = 7.0/4.0
- Three from racial group 2 with Weight2 = 2.0/3.0
- Three from racial group 3 with Weight2 = 1.0/3.0

Final weights are obtained as follows:

```
Final_wt = Weight3*Weight2*10.0.
```

Multiplication of the weights by a factor of 10.0 was done to illustrate that a constant scaling of the weights does not affect parameter estimates, standard error estimates or the chi-square goodness of fit statistic value.
The data were simulated according to the following model.

\[ Score_j = a_{0i} + a_i j + e_j, \quad j = 1, 2 \ldots 6 \]

In the model, \( t_j = (j - 1) \) and \( i \) denotes student number \( i \). In simulating the data set, it was assumed that

\[
a_{0i} = \alpha_0 + \gamma_1 \text{Lang}1 + \gamma_2 \text{Lang}2 + u_{0i} \\
a_{ii} = \alpha_i + u_{ii}
\]

where

\[
\begin{pmatrix}
\alpha_0 \\
\alpha_1
\end{pmatrix} = \begin{pmatrix}
1.0 \\
0.5
\end{pmatrix}
\]

\[
\psi_{11} = \text{var}(u_{0i}) = 2 \\
\psi_{21} = \text{cov}(u_{0i}, u_{ii}) = 0.6 \\
\psi_{22} = \text{var}(u_{ii}) = 0.4
\]

\[ e_j \sim N(0, \sigma^2), \quad \sigma^2 = 1 \]

and

\[
\begin{pmatrix}
\gamma_1 \\
\gamma_2
\end{pmatrix} = \begin{pmatrix}
0.5 \\
-1.0
\end{pmatrix}
\]

The first 10 records of the data set are shown below.

Note that even-numbered cases have missing values on Score5 and Score6. This is indicated with a missing value code of -9.00.
3.3.2.2 The model

The conceptual path diagram for the model is shown below. The paths from the latent variables intercept and time to the dependent variables Score\(_1\) to Score\(_6\) are shown in gray to indicate that the corresponding coefficients are fixed values.

The conceptual path diagram for the structural part of the model is given below and indicates that allowance is made for the latent variables to be correlated.

Mathematical Model

Measurement model

The measurement model for the latent variables may be expressed as

\[
\begin{bmatrix}
  y \\
  x
\end{bmatrix} = \begin{bmatrix}
  \Lambda_y & 0 \\
  0 & \Lambda_x
\end{bmatrix} \begin{bmatrix}
  \eta \\
  \xi
\end{bmatrix} + \begin{bmatrix}
  \varepsilon \\
  \delta
\end{bmatrix}
\]
where \( y = [ \text{Score1} \ \text{Score2} \ \ldots \ \text{Score6} ]' \), \( x = [ \text{Lang1} \ \text{Lang2} ]' \), \( \eta = [ \text{intcept} \ \text{time} ]' \), \( \xi = [ \text{Lang1} \ \text{Lang2} ]' \),
\( \delta = [ \delta_1 \ \delta_2 ]' \), \( \varepsilon = [ \varepsilon_1 \ldots \varepsilon_6 ] \) and \( \text{Cov}(\varepsilon) = \sigma^2 \mathbf{I} \). Also

\[
\Lambda_y = \begin{bmatrix}
1 & 0 \\
1 & 1 \\
1 & 2 \\
1 & 3 \\
1 & 4 \\
1 & 5
\end{bmatrix},
\]

\[
\text{Cov}(\xi) = \begin{bmatrix}
\phi_{11} & \phi_{21} \\
\phi_{21} & \phi_{22}
\end{bmatrix},
\]

\[
E(\xi) = \begin{bmatrix}
\kappa_1 \\
\kappa_2
\end{bmatrix},
\]

\[
\Lambda_x = \begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix},
\]

and \( \text{Cov}(\delta) = 0 \).

**Structural equation model**

The structural equation model for the latent variables intcept and time is given by \( (B = 0) \)

\[
\eta = \Gamma \xi + \zeta
\]

where \( \zeta = [ \zeta_1 \ \zeta_2 ]' \), with

\[
\text{Cov}(\zeta) = \begin{bmatrix}
\psi_{11} & \psi_{21} \\
\psi_{21} & \psi_{22}
\end{bmatrix},
\]

\[
E(\eta) = \begin{bmatrix}
\alpha_1 \\
\alpha_2
\end{bmatrix},
\]

and

57
where $\gamma_1$ and $\gamma_2$ denote unknown regression weights. The thirteen unknown model parameters are $\gamma_1$, $\gamma_2$, $\psi_{11}$, $\psi_{12}$, $\psi_{22}$, $\alpha_1$, $\alpha_2$, $\sigma_e^2$, $\phi_{11}$, $\phi_{21}$, $\phi_{22}$, $\kappa_1$, and $\kappa_2$.

The unrestricted model has $36 + 8 = 44$ parameters, since there are 8 observed variables and the number of non-duplicated elements of a covariance matrix of order 8 is 36. Therefore, the number of degrees of freedom equals $44 - 13 = 31$.

SIMPLIS syntax for the model is shown next. Note that $5.0 \times \text{time}$, for example, indicates that the coefficient for the time→ Score6 path is fixed at a value of 5.0.

An experienced LISREL user may prefer to type the SIMPLIS commands to fit a specific model. However, we provide an outline of the steps required to build the syntax by drawing a path diagram. With more experience, users will find many shortcuts in developing the syntax. For example, it may be easier to type in the $1.0^*$, $2.0^*$, etc. values after syntax has been generated than to fix and set each path to a specific value.
3.3.3 Implementation of sampling weights in a linear growth curve model

3.3.3.1 Describing the data

From the main menu bar, select the Data, Survey Design option and add the variable Final_wt in the Design weight: box. Click OK when done.

Next use the Data, Define Variables option to select the variables Score5 and Score6, and then click the Missing Values button to invoke the Missing Values dialog box. Enter -9.0 as shown below. Click OK, and then use the File, Save option to ensure that these changes are contained in the LSF file.
3.3.3.2 Setting up the analysis

To generate the SIMPLIS commands interactively, we proceed as follows. Using the **File, New** option, select the **Path Diagram** option from the **New** dialog box.
Save the new path diagram in the **MISSINGEX** folder as **Survey1.pth**.

Click **Save** when done. From the **Setup** menu, select **Title and Comments**.
This action loads the **Title and Comments** dialog box shown below. Enter a title and any (optional) comments as shown below.

![Title and Comments dialog box]

Click **Next** to proceed to the **Groups** dialog and, since this is a single-group example, click **Next** again to activate the **Labels** dialog shown below. To add a list of observed variables, click the **Add/Read Variables** button below the **Observed Variables** list box.

![Labels dialog box]

From the **Add/Read Variables** dialog box, click the **Read from file**: radio button and select **Lisrel System File**. Next, use the **Browse** button to locate and select **surveysem.LSF**.
Click the OK button once the desired LSF is selected. The observed variable names will be displayed in the Labels dialog box. To add a list of latent variables, click the Add Latent Variables button and type the names of the latent variables one at a time.
Click **OK** when the latent variables intercept and time are entered. The left hand side of the path diagram window should display the observed and latent variables. If not, select the **View, Toolbars** option and from the drop-down list **Select Variables**. Click on the check boxes on the right hand side of the variable names to define Score1 to Score6 as Y (dependent) variables and intercept and time as Eta (endogenous latent) variables.

We start drawing the path diagram by dragging the names Score1 to Score6 to the path diagram window. Next, drag intercept and time to the middle of the path diagram window. A variable is dragged to the path diagram window by left-clicking on the variable name and then moving it with the left mouse button held down. Finally, drag Lang1, Lang2 and CONST to the left of intercept and time. Click the one-sided arrow on the drawing bar and connect arrows from intercept to the variables Score1, ..., Score6.
With the left mouse button down, start in the ellipse and drag the arrow to within a rectangle representing one of the Score variables before releasing the mouse button. Unselect the arrow by clicking on the square on the drawing bar. Once this is done, move the mouse pointer to each arrow and right-click. From the pop-up menu, select Fix. Repeat this for each path from intercept to a Score value.

Once all the paths are fixed, start with the path from intercept to Score1. Right-click on the arrow, and select the Set Value option from the pop-up menu.
Change the value of each path to 1.0.

Repeat the above procedure by drawing paths from \textit{time} to \textit{Score2}, \textit{Score3}, ..., \textit{Score6}. Fix these paths and set the path coefficients to 1.0, 2.0, 3.0, 4.0, and 5.0 respectively, as shown below.
The path diagram is completed by drawing arrows from \(\text{Lang1}\), \(\text{Lang2}\) and the \text{SIMPLIS} variable \text{CONST} to \text{intcept} and \text{time} as shown. Once this is done, select the two-sided arrow (error covariance or factor correlation) to add a covariance path between \(\text{Lang1}\) and \(\text{Lang2}\).

![Path Diagram](image.png)

To draw this path, left-click on the horizontal link between 0.00 and the \(\text{Lang1}\) rectangle. Drag the path to the line connecting 0.00 and \(\text{Lang2}\) before releasing the mouse button.

![Path Diagram](image2.png)
To build the corresponding SIMPLIS syntax, select Setup, Build SIMPLIS Syntax from the main menu.

The syntax shown below is generated.

In our model, we assume that the coefficient of the latent variable time is not influenced by Lang1 or Lang2 and we change the syntax by adding 0.00* in front of Lang1 and Lang2 as shown below.
Since it is assumed that the error variances associated with the dependent variables are homogeneous, we manually add the command

\[ \text{Equal Error Variances: Score1 – Score6} \]

In addition, it is assumed that \( \text{intcept} \) and \( \text{time} \) are correlated. The equivalent SIMPLIS command is also typed in, as shown in the syntax file.

Once these modifications to the command file are completed, click the **Run LISREL** icon to obtain a path diagram.

### 3.3.3.3 Discussion of results – implementation of sampling weights in a linear growth curve model

The path diagram shows the parameter estimates and \( \chi^2 \) goodness of fit statistic (\( \chi^2 = 57.15, \text{df} = 31 \)) under the assumption that stratification and selection of clusters do not affect standard errors or model fit.
In the display below, the structural part of the model is shown. The numeric values are the $t$-values ($t$: parameter estimate/ std. error) for each path. For example; the $t$-value for $\beta_1 = 0.36$ equals 1.99.
3.3.4 Incorporating stratum and cluster variables in the model

Next, select the **Survey Design** option from the **Data** menu and add County and School respectively as stratification and cluster variables. Click **OK** and then select the **File, Save** option.

Once this is done, make `survey1.spj` the active window before clicking the **Run Lisrel** icon.
3.3.4.1 Discussion of results – Incorporating stratum and cluster variables in the model

From the path diagram below we note that the $\chi^2$ statistic has increased from 57.15 to 62.76 and that the $t$-value corresponding to $\hat{\gamma}_1$ decreased from 1.99 to 1.68. In the latter case, the $t$-value indicates a non-significant coefficient for the Lang1, intercept path.

![Path Diagram](image)

Chi-Square=62.76, df=31, P-value=0.00063, RMSEA=0.046

Finally, selected parts of the LISREL output are shown below.
The 95% confidence intervals for the estimated parameters show that the population parameter values are all contained in the respective intervals.
The $\chi^2$ goodness-of-fit statistic (see below) equals 62.757. This value is obtained as the difference between the deviance values for the saturated and fitted models, multiplied by the chi-squared scale factor. That is $(9484.714 - 9408.501) \times 0.82344$. The RMSEA value of 0.0462 indicates that the model provides a reasonable fit to the data.
3.4 Evaluation

3.4.1 Simulation study based on a linear growth curve model

3.4.1.1 Introduction

A simulation study based on a linear growth model for continuous outcomes is discussed in this section. A structural equation model, mathematically equivalent to a multilevel model, is fitted to the data.

3.4.1.2 The data

An unbalanced design, consisting of 500 univariate observations that are clustered within 100 level-two units, was used. Half of the level two units have four observations and the other half have six observations. The times of the observations are equally spaced starting at 0 and ending with 3 for the clusters with 4 observations and ending with 5 for the clusters with 6 observations. The linear growth model has random intercept and slope coefficients.

3.4.1.3 The model

A path diagram of the model is shown below.
The LISREL model is

\[ x = \Lambda_x \xi + \delta \]

where

\[ \Lambda_x = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix}, \]

\[ E(\xi) = \begin{bmatrix} \kappa_1 \\ \kappa_2 \end{bmatrix}, \quad Cov(\xi) = \begin{bmatrix} \phi_{11} & \phi_{21} \\ \phi_{21} & \phi_{22} \end{bmatrix}, \]

and \( Cov(\delta) = \sigma^2 I \).

### 3.4.1.4 Setting up the analysis

The SIMPLIS syntax for fitting this growth model is as follows.
The command `intercept time = CONST` indicates the estimation of the population intercept and slope coefficients. The command `Equal Error Variances: Y1 – Y6` specifies a homogeneous error variance term on level-1 of the model.

### 3.4.1.5 Discussion of results

Results of the simulation study are summarized in Tables 3.1 and 3.2.

**Table 3.1: Bias and Coverage in LISREL and Mplus**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Value</th>
<th>Bias LISREL</th>
<th>Bias Mplus</th>
<th>Coverage LISREL</th>
<th>Coverage Mplus</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>0.5</td>
<td>0.017</td>
<td>0.017</td>
<td>0.908</td>
<td>0.908</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.1</td>
<td>0.002</td>
<td>0.002</td>
<td>0.938</td>
<td>0.942</td>
</tr>
<tr>
<td>$\phi_{11}$</td>
<td>1</td>
<td>-0.024</td>
<td>-0.024</td>
<td>0.845</td>
<td>0.848</td>
</tr>
<tr>
<td>$\phi_{22}$</td>
<td>0.2</td>
<td>-0.006</td>
<td>-0.006</td>
<td>0.892</td>
<td>0.902</td>
</tr>
<tr>
<td>$\phi_{21}$</td>
<td>0.3</td>
<td>-0.006</td>
<td>-0.006</td>
<td>0.936</td>
<td>0.940</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>1</td>
<td>-0.008</td>
<td>-0.008</td>
<td>0.908</td>
<td>0.910</td>
</tr>
</tbody>
</table>

The bias and coverage produced by LISREL and Mplus are virtually identical.

As part of the simulation study, the unadjusted and adjusted $\chi^2$ goodness-of-fit statistics for each of the 500 simulations were computed. The adjusted $\chi^2$ was obtained by the multiplication of the unadjusted
\( \chi^2 \) with the \( \chi^2 \) scale factor, as described in Section 3.6. The mean values and rejection rates of these statistics for the 500 samples are given in Table 3.2. The rejection rate (expressed as a percentage) denotes the number of times the \( \chi^2 \) statistic exceeded \( \chi^2_{21,0.05} \). The degrees of freedom, 21, is obtained as the number of non-duplicated elements of the covariance matrix plus the number of means minus the number of parameters estimated. Therefore \( df = 21 + 6 - 6 \).

Table 3.2: Mean values and rejection rates of \( \chi^2 \) goodness-of-fit statistics

<table>
<thead>
<tr>
<th>Unadjusted ( \chi^2 )</th>
<th>Adjusted ( \chi^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>32.98</td>
<td>22.49</td>
</tr>
<tr>
<td>32 %</td>
<td>9 %</td>
</tr>
</tbody>
</table>

The expected mean for a \( \chi^2_{21} \) random variate is 21. The mean for the adjusted \( \chi^2 \), which is higher than the expected mean, explains the rejection rate being higher than 5 %. This result implies that more research on the correction factor of the \( \chi^2 \) test statistic under complex sampling is indicated.

3.4.2 Latent curve analysis with main and interaction effects

3.4.2.1 Introduction

Curran, Bauer & Willoughby (2004) considered the testing of main effects and interactions in latent curve analysis. Their goal was to illustrate that classic techniques, as applied in multiple regression, can be generalized to the case of latent curve analyses. As part of the paper, an example was used to illustrate the testing of a categorical by continuous interaction in an unbalanced latent curve model with missing data over time. In this section, the same model is fitted, with and without sampling weights, in order to evaluate the impact of ignoring weights on an analysis: an analysis option not available to the authors of the paper at the time of publication.

3.4.2.2 The data

The example in the paper was based on data from the National Longitudinal Study of Youth. Specific details on the selection of the sample can be found in Curran (1997). The sample consists of information on 405 children at 4 occasions. At the start of the study, children in the sample were between 6 and 8 years old. Information is not available for all children on all occasions: while 405 were interviewed initially, the second, third, and fourth occasions provided information on 374, 297, and 294 children respectively. Only 221 children were interviewed on all four occasions. As such, the participant attrition over time, combined with the variability in age at the start of the study and the fact that measurement occasions were approximately two years apart, makes this an example of an unbalanced design with missing data.
In this section, the same data are used. Two analyses will be performed: a multilevel and a SEM analysis, the latter to verify the validity of the comparison of our results with that of Curran, Bauer & Willoughby (2004). In addition, models will be fitted with and without sampling weights.

The data were used in different formats for the structural equation and multilevel models. A short description of each data format is given below.

**Structural equation modeling**

A few of the variables in the data set *curran_NLSY_subset.LSF* are shown below for the first 10 observations.

![Table of variables](image)

The emotional support at home and the level of antisocial behavior exhibited by these children were of special interest. The authors focused on three questions of interest: the form of the mean developmental trajectory of antisocial behavior over time, the possibility of meaningful individual variability in trajectories around these mean values, and the possible effect of interaction between the gender of a child and the level of emotional support on antisocial behavior.

The following variables included in the LSF were selected from the survey data:

- **MOM_ID**: This variable represents the identification number of the mother and serves as grouping variable for all measurements for a specific child. There are 405 mothers included in this subset of the NLSY data.
- **MOM_WT**: The sampling weight for each mother.
- **antiy1 – antiy10**: A measure of antisocial behavior in the child. For each of these variables, a continuous measure representing the sum of six items assessing child antisocial behavior over the previous 3 months was created with values ranging between 0 and 12, where a high value would indicate a higher level of antisocial behavior.
- **genfemo**: The gender of the child, coded “0” for a female, and “1” for a male.
- **home_emo**: A measure of emotional support of the child in the home. This continuous measure, ranged from 0 to 13 with higher values reflecting higher levels of support, was
measured at the first measurement occasion. It is centered around the mean level of emotional support.

- **home_cog**: A measure of cognitive stimulation, based on a summation of 14 dichotomously scored items reported by the mother.
- **genxemo**: A variable intended to represent the interaction between a child’s gender and level of emotional support: \( \text{genxemo} = \text{genfemo} \times \text{home}_\text{emo} \).

### Multilevel modeling

For the multilevel analysis, the LSF **curran_mlev.LSF** was used as basis of the analysis. Data on all the variables used in this model are shown below for the first 10 respondents. Note that, in contrast to the LSF used for the SEM model, antisocial behavior is now represented by a single variable containing the stacked measurements over the 4 measurement occasions. For the first child, for example, the values 2, 1, 0, and 2 respectively were observed at the measurement occasions, where the latter is indicated by the variable tim.

<table>
<thead>
<tr>
<th>MOM_ID</th>
<th>tim</th>
<th>Antiy</th>
<th>MOM_Wt</th>
<th>genfemo</th>
<th>homo_emo</th>
<th>home_cog</th>
<th>genxemo</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>2.00</td>
<td>498090.00</td>
<td>0.00</td>
<td>-1.98</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
<td>3.00</td>
<td>498090.00</td>
<td>0.00</td>
<td>-1.98</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.00</td>
<td>5.00</td>
<td>159980.00</td>
<td>0.00</td>
<td>-1.08</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.00</td>
<td>7.00</td>
<td>159980.00</td>
<td>0.00</td>
<td>-1.08</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.00</td>
<td>2.00</td>
<td>596290.00</td>
<td>1.00</td>
<td>-2.00</td>
<td>-1.00</td>
<td>-0.20</td>
</tr>
<tr>
<td>6</td>
<td>2.00</td>
<td>0.00</td>
<td>159980.00</td>
<td>0.00</td>
<td>-0.20</td>
<td>-1.00</td>
<td>-0.20</td>
</tr>
<tr>
<td>7</td>
<td>3.00</td>
<td>1.00</td>
<td>159980.00</td>
<td>1.00</td>
<td>1.79</td>
<td>1.11</td>
<td>1.79</td>
</tr>
<tr>
<td>8</td>
<td>2.00</td>
<td>0.00</td>
<td>159980.00</td>
<td>1.00</td>
<td>1.79</td>
<td>1.11</td>
<td>1.79</td>
</tr>
<tr>
<td>9</td>
<td>3.00</td>
<td>0.00</td>
<td>596290.00</td>
<td>0.00</td>
<td>1.11</td>
<td>1.79</td>
<td>1.79</td>
</tr>
<tr>
<td>10</td>
<td>3.00</td>
<td>0.00</td>
<td>596290.00</td>
<td>0.00</td>
<td>1.11</td>
<td>1.79</td>
<td>1.79</td>
</tr>
</tbody>
</table>

The following variables were used in the multilevel analysis:

- **Mom_ID**: This variable represents the identification number of the mother and serves as grouping variable for all measurements for a specific child. There are 405 mothers included in this subset of the NLSY data.
- **tim**: This variable indicates the time of measurement, and varies from 0 to 9.
- **Antiy**: A measure of antisocial behavior in the child at a given measurement occasion. This continuous measure, representing the sum of six items assessing child antisocial behavior over the previous 3 months, was created with values ranging between 0 and 12, where a high value would indicate a higher level of antisocial behavior.
- **Mom_Wt**: The sampling weight for each mother.
- **genfemo**: The gender of the child, coded “0” for a female, and “1” for a male.
- **home_emo**: A measure of emotional support of the child in the home. This continuous measure ranged from 0 to 13, with higher values reflecting higher levels of support, was measured at the first measurement occasion. It is centered around the mean level of emotional support.
- **home_cog**: A measure of cognitive stimulation, based on a summation of 14 dichotomously scored items reported by the mother.
- **genxemo**: A variable intended to represent the interaction between a child’s gender and level of emotional support: $\text{genxemo} = \text{genfemo} \times \text{home_emo}$.

### 3.4.2.3 The model

Curran, Bauer & Willoughby (2004) shows how a structural equation model-based latent curve analysis and a hierarchical linear model for these data can be formulated to produce equivalent results. They point out, however, that there are subtle but important differences in both model estimation and interpretation due to the way in which time is incorporated into the model. These differences are of particular importance in the case of conditional growth models, where one or more exogenous variables predict the random growth curve parameters. Main effects of the random trajectories imply that exogenous variables interact with time in the prediction of repeated measures for both cases. While both predictors and time are used as exogenous variables in the hierarchical linear model, the interaction between time and any predictor is explicitly modeled as a cross-level interaction. The latent curve analysis does not use time as a variable as such. Instead, it is incorporated into the model via the factor loading matrix. In this section, the two models and data sets constructed for use in the analyses will further illustrate this difference. To accommodate the differences in models fitted and data sets used, the structural equation model and the multilevel, or hierarchical linear, model, will be discussed separately in the rest of this section.

**Structural equation model**

We first consider the structural equation model. The model shown below corresponds to Figure 2 in Curran, Bauer & Willoughby (2004), and represents cohort-sequential conditional linear latent curve model with 10 time points, regressed on the main effect of gender, the main effect of emotion, and the interaction between gender and emotion. The variables intercept and slope represent the latent intercept and latent slope of the trajectory respectively.
In the Y part of the model, we include the dependent variables \( \text{antiy1} \) to \( \text{antiy10} \). It is assumed that \( \text{antiy1} \) to \( \text{antiy10} \) are indicators of the endogenous (ETA) latent variables intercept and slope.

The covariates \( \text{genfemo} \), \( \text{home_emo} \), and \( \text{genxemo} \) are also assumed to have relationships with both the intercept and the slope of the trajectory and form the X part of the model. Finally, we allow the intercept and slope variables to be correlated. This path cannot be seen on the basic path diagram, which is the type of diagram used here (to view this path, select the **Structural Model** option from the **Models**: drop-down list in the PTH window).

**Mathematical model**

The LISREL model consists of a measurement and structural part.

**Measurement model**

The measurement model may be expressed as

\[
\begin{bmatrix}
\text{y} \\
\text{x}
\end{bmatrix} =
\begin{bmatrix}
\Lambda_y & 0 \\
0 & \Lambda_x
\end{bmatrix}
\begin{bmatrix}
\eta \\
\xi
\end{bmatrix} +
\begin{bmatrix}
\varepsilon \\
\delta
\end{bmatrix}
\]
where \( y = \begin{bmatrix} \text{antiy1} & \text{antiy2} & \ldots & \text{antiy10} \end{bmatrix}, \quad x = \begin{bmatrix} \text{genfemo} & \text{hom_emo} & \text{genxemo} \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \ldots & \varepsilon_{10} \end{bmatrix}, \)
\( \delta = \begin{bmatrix} \delta_1 & \delta_2 & \delta_3 \end{bmatrix}, \quad \xi = \begin{bmatrix} \text{genfemo} & \text{hom_emo} & \text{genxemo} \end{bmatrix}, \)

\[ \eta = \begin{bmatrix} \text{intcept} \\ \text{slope} \end{bmatrix} \]

and

\[ \Lambda_y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{bmatrix}. \]

\( \text{Cov}(\varepsilon) \) is a diagonal matrix with diagonal elements \( \text{var}(\varepsilon_1), \text{var}(\varepsilon_2), \ldots, \text{var}(\varepsilon_{10}) \) where we constrained these elements to be equal, while \( \Lambda_x \) is a 3×3 identity matrix and

\[ \text{Cov}(\xi) = \Phi = \begin{bmatrix} \phi_{11} & \phi_{21} & \phi_{31} \\ \phi_{21} & \phi_{22} & \phi_{32} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{bmatrix}. \]

**Structural equation model**

The structural equation model for the latent variables intercept and slope is given by (\( B = 0 \))

\[ \eta = \Gamma \xi + \zeta \]

where \( \zeta = \begin{bmatrix} \zeta_1 & \zeta_2 \end{bmatrix} \), with

\[ \text{Cov}(\zeta) = \begin{bmatrix} \psi_{11} & \psi_{21} \\ \psi_{21} & \psi_{22} \end{bmatrix}, \]

\[ E(\zeta) = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}, \]
and

\[ \mathbf{\Gamma} = \begin{bmatrix} \gamma_1 & \gamma_2 & \gamma_3 \\ \gamma_4 & \gamma_5 & \gamma_6 \end{bmatrix}. \]

The unknown model parameters are therefore \( \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6, \alpha_1, \alpha_2, \phi_1, \phi_2, \ldots, \phi_3, \psi_{11}, \psi_{12} \), \( \psi_{22} \), and \( \text{var}(e_i) \).

**Multilevel model**

A general two-level model for a response variable \( y \) depending on a set of \( r \) predictors \( x_1, x_2, \ldots, x_r \) can be written in the form

\[ y_{ij} = \mathbf{x}_{(f)ij}' \mathbf{\beta} + \mathbf{x}_{(2)i}' \mathbf{u}_i + \mathbf{x}_{(1)ij}' \mathbf{e}_j. \]

Where \( i = 1, 2, \ldots, N \) denotes the level-2 units, and \( j = 1, 2, \ldots, n_i \) the level-1 units. Thus \( y_{ij} \) represents the response of individual \( j \), nested within level-2 unit \( i \). The model shown here consists of a fixed and a random part. The fixed part of the model is represented by the vector product \( \mathbf{x}_{(f)ij}' \mathbf{\beta} \), where \( \mathbf{x}_{(f)ij} \) is a typical row of the design matrix of the fixed part of the model with, as elements, a subset of the \( r \) predictors. The vector \( \mathbf{\beta} \) contains the fixed, but unknown parameters to be estimated. The vector products \( \mathbf{x}_{(2)i}' \mathbf{u}_i \) and \( \mathbf{x}_{(1)ij}' \mathbf{e}_j \) denote the random part of the model at levels 2 and 1 respectively. For example, \( \mathbf{x}_{(2)i}' \) represents a typical row of the design matrix of the random part at level-2, and \( \mathbf{u}_i \) the vector of random level-2 coefficients to be estimated. The product \( \mathbf{x}_{(1)ij}' \mathbf{e}_j \) serves the same purpose at level-1. It is assumed that \( \mathbf{u}_i, \mathbf{u}_2, \ldots, \mathbf{u}_N \) are assumed i.i.d., with mean vector \( \mathbf{0} \) and covariance matrix \( \Phi_{(2)} \), and \( \mathbf{e}_{i1}, \mathbf{e}_{i2}, \ldots, \mathbf{e}_{im} \) are assumed i.i.d., with mean vector \( \mathbf{0} \) and covariance matrix \( \Phi_{(1)} \).

Within this hierarchical framework, the model fitted to the data uses the participant's gender, level of emotional support at home, and the interaction between these variables to predict the variability in intercept and slope over time of antisocial behavior trajectories.

\[
\text{antiy}_{ij} = \beta_0 + \beta_1 \ast \text{genfemo}_{ij} + \beta_2 \ast \text{home emo}_{ij} + \beta_3 \ast \text{genxemo}_{ij} + \beta_4 \ast \text{tim}_{ij} + \beta_5 \ast (\text{genfemo}_{ij})(\text{tim}_{ij}) + \beta_6 \ast (\text{home emo}_{ij})(\text{tim}_{ij}) + \beta_7 \ast (\text{genxemo}_{ij})(\text{tim}_{ij}) + u_{i0} + u_{i1} \ast \text{tim}_{ij} + e_{ij}
\]
where $\beta_0$ denotes the average expected level of antisocial behavior for a female child at the first measurement occasion with a score of 0 on the measure of emotional support at home. The coefficients $\beta_1, \beta_2, \ldots, \beta_7$ are the estimated coefficients associated with the fixed part of the model which contains the predictor variables genfemo, home_emo, and the interaction term genxemo. The random part of the model is represented by $u_{i0}$, $u_{i0} \times \text{tim}_i$, and $e_{i}$, which denote the variation in average level of antisocial behavior between children, in slope over measurements occasions, and between measurements taken at different occasions, where the occasions form the lowest level of the hierarchy.

### 3.4.2.4 Setting up the analysis

#### 3.4.2.4.1 Structural equation model

The SIMPLIS syntax for the model is shown below. Note that 5.0*slope, for example, indicates that the coefficient for the slope→antiy6 path is fixed at a value of 5.0.

![SIMPLIS Syntax Example](image.png)

An experienced LISREL user may prefer to type the SIMPLIS commands to fit a specific model. Alternatively, the syntax can be created by drawing a path diagram. This is done in the same way as shown in Section 3.4.2. To add a weight variable, as is the case in the second of the structural equation models considered here, select the **Data, Survey Design** option from the main menu bar.
and enter the variable MOM_WT in the **Design weight** field as shown below.

To run the model, click the **Run LISREL** icon button on the main menu bar.

### 3.4.2.4.2 Multilevel model

Specifying the multilevel model is straightforward, and proceeds as discussed in the *Multilevel Examples Guide*. Briefly, the level-2 ID is identified as Mom_ID (see **Identification Variables** dialog box below), the outcome is antiy, and the fixed part of the model consists of the variables genfemo, home_emo, genxemo and tim as shown in the **Select Response and Fixed Variables** dialog box below. The weight variable
Mom_Wt used in the second of the multilevel analyses discussed here is entered in the **Weight Variables** dialog box.
Note that, in the Select Response and Fixed Variables dialog box shown, the required interaction between tim and the three variables genfemo, home_emo, genxemo is not included – this will be added manually to the syntax file created by the GUI.
To estimate both a random intercept and a random slope, add the variable tim to the Random Level-2 field on the Random Variables dialog box as shown below.
The syntax generated via the **Finish** button on the **Random Variables** dialog box is shown below:

![Random Variables dialog box]

To build syntax, proceed to the Random Variables screen and click the **Finish** button.

Finally, type the additional interaction terms `tim*genfemo tim*home_emo tim*genxemo` into the syntax file to obtain the final syntax as shown below. The analysis is started by clicking the **Run Prelis** icon button on the main menu bar.

![curran_mlev1.PRL]

```plaintext
OPTIONS OLS=YES CONVERGE=0.001000 MAXITER=15 OUTPUT=STANDARD:
TITLE=Curran 2004 paper using weights;
SY='C:\LISREL9\Examples\MISSINGEX\curran_mlev.LSF';
ID2=MOM_ID;
WEIGHT1=MOM WT; RESPONSE=Anty;
FIXED=intercept tim genfemo home_emo genxemo;
RANDOM1=intercept tim;
RANDOM2=intercept tim;
```
3.4.2.5 Discussion of results

The results of both the SEM and multilevel models are summarized in Table 3.3. The table also contains the results from Curran, Bauer & Willoughby (2004). These analyses did not take the sampling weight Mom_Wt into account. The results of the weighted analysis, in which this variable was incorporated into the estimation procedure, are reported in Table 3.6.

Table 3.3: Unweighted analyses: comparison of results

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimates</th>
<th>Standard errors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CBW paper</td>
<td>Multilevel</td>
</tr>
<tr>
<td>genfemo ($\gamma_1$)</td>
<td>0.829</td>
<td>0.829</td>
</tr>
<tr>
<td>home_emo ($\gamma_2$)</td>
<td>-0.194</td>
<td>-0.194</td>
</tr>
<tr>
<td>genxemo ($\gamma_3$)</td>
<td>0.044</td>
<td>0.044</td>
</tr>
<tr>
<td>INTCPT ($\alpha_0$)</td>
<td>1.217</td>
<td>1.217</td>
</tr>
<tr>
<td>tim $\times$ genfemo ($\gamma_4$)</td>
<td>0.013</td>
<td>0.013</td>
</tr>
<tr>
<td>tim $\times$ home_emo ($\gamma_5$)</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td>tim $\times$ genfemo $\times$ home_emo ($\gamma_6$)</td>
<td>-0.029</td>
<td>-0.029</td>
</tr>
<tr>
<td>tim ($\alpha_1$)</td>
<td>0.066</td>
<td>0.066</td>
</tr>
<tr>
<td>Var(intcept) ($\phi_{11}$)</td>
<td>*</td>
<td>0.669</td>
</tr>
<tr>
<td>Var(time slope) ($\phi_{22}$)</td>
<td>*</td>
<td>0.019</td>
</tr>
<tr>
<td>Cov(intcept,tim) ($\phi_{21}$)</td>
<td>*</td>
<td>0.076</td>
</tr>
<tr>
<td>$\sigma^2$ (var($\varepsilon$))</td>
<td>*</td>
<td>1.758</td>
</tr>
</tbody>
</table>

* Not reported in the Curran et. al. paper
Table 3.4: Weighted analyses: comparison of results

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimates Multilevel</th>
<th>Standard errors Multilevel</th>
<th>SEM</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>genfemo (γ₁)</td>
<td>0.901</td>
<td>0.201</td>
<td>0.205</td>
<td></td>
</tr>
<tr>
<td>home_emo (γ₂)</td>
<td>-0.244</td>
<td>0.021</td>
<td>0.058</td>
<td></td>
</tr>
<tr>
<td>genxemo (γ₃)</td>
<td>0.122</td>
<td>0.112</td>
<td>0.115</td>
<td></td>
</tr>
<tr>
<td>INTCPT (α₀)</td>
<td>1.203</td>
<td>0.120</td>
<td>0.115</td>
<td></td>
</tr>
<tr>
<td>tim × genfemo (γ₄)</td>
<td>0.024</td>
<td>0.047</td>
<td>0.048</td>
<td></td>
</tr>
<tr>
<td>tim × home_emo (γ₅)</td>
<td>0.014</td>
<td>0.015</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td>tim × genfemo × home_emo (γ₆)</td>
<td>-0.020</td>
<td>0.026</td>
<td>0.038</td>
<td></td>
</tr>
<tr>
<td>tim (α₁)</td>
<td>0.073</td>
<td>0.030</td>
<td>0.030</td>
<td></td>
</tr>
<tr>
<td>Var(intcept) (φ₁₁)</td>
<td>0.487</td>
<td>0.317</td>
<td>0.330</td>
<td></td>
</tr>
<tr>
<td>Var(time slope) (φ₂₂)</td>
<td>0.230</td>
<td>0.014</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td>Cov(intcept, tim) (φ₂₁)</td>
<td>0.091</td>
<td>0.060</td>
<td>0.063</td>
<td></td>
</tr>
<tr>
<td>σ² (var(ε))</td>
<td>1.966</td>
<td>0.234</td>
<td>0.217</td>
<td></td>
</tr>
</tbody>
</table>

The goodness-of-fit of the models fitted can also be compared. For the weighted structural equations model, the following path diagram was obtained.

From the path diagram, \( \chi^2 = 249.52 \), with 83 degrees of freedom. The corresponding \( \chi^2 \)-statistic value for the unweighted model is 107.2978, with 83 degrees of freedom. For the analyses that included the weight variable, there are differences in parameter estimates if we compare the multilevel model results with those of the structural equation model. These differences are most evident in the covariance matrix of the latent variables. LISREL produced a warning message that this matrix is not positive definite. A matrix that is not positive definite can have a large impact on the estimated chi-square value. On further examination, it was found that the variable antiy10 contained only 8 non-missing observations. Refitting of the models, using the first 9 variables antiy1 – antiy9, showed that the multilevel and SEM results are identical in this case. The \( \chi^2 \)-statistic in this case is equal to 155.493.
3.4.3 Replicate weights

3.4.3.1 Introduction

Survey data sets often include a column $W_0$ of design weights and additional columns $W_1, W_2, \ldots, W_{R-1}$ of replicate weights. Typically, a researcher may repeatedly fit the same model to the data by working through the sequence of weight variables $W_0, W_1, \ldots, W_{R-1}$. Means of the $R$ sets of parameter estimates and their standard errors may subsequently be computed to obtain more accurate parameter and standard error estimates. In this section, we illustrate how to use LISREL in the case of replicate weights.

3.4.3.2 The data

The data set used here is `missingex\replicwts.lsf`. A few of the variables are shown below for the first 10 observations in the data set.
The contents of the LSF are obtained by selecting the **Statistics, Data Screening** option. A portion of the output is shown below.

The following variables included in the LSF were selected from the survey data:

- **CENREG**: This variable indicates the census region and has four categories, these being "Northeast," "Midwest," "South," and "West" respectively.
- **FACTYPE**: The facility treatment type has four categories, too, representing facilities with "residential treatment", "outpatient methadone treatment", "outpatient non-methadone treatment", and "more than one type of treatment" respectively.
- **ALCEU**: An indicator variable with value "1" if the respondent has ever used alcohol, and "0" otherwise.
- **COCEU**: An indicator variable with value "1" if the respondent has ever used cocaine, and "0" otherwise.
- **MAREU**: An indicator variable with value "1" if the respondent has ever used marijuana, and "0" otherwise.
- **AGE**: This variable denotes age at admission to a facility.
- **GENGER**: The respondent's gender is denoted by this indicator variable that assumes a value of "1" for female respondents.
- **RACE_D**: The original variable RACE recoded so that '1' denotes white and "0" other ethnic groups.
- **DEPR**: This indicator variable is coded "1" if the respondent is depressed, and "0" otherwise.
- **EDU**: A categorical variable representing the respondent's level of education at admission. It has 5 categories, these being (from 1 to 5) "less than 8 years", "8 – 11 years or less than High School graduate", "High School graduate / GED", "some college", and "college graduate / postgraduate".
- **JAILR**: This indicator variable indicates whether the respondent had a prison or jail record prior to admission.
- **NUMTE**: A count variable, indicating the total number of treatment episodes prior to admission.
- **A2TWA0 – A2TWA78**: These variables are abstract final full sample weights. A more complete description follows below.

Variance estimation frequently relies on one of two techniques: Taylor series linearization or replication weights. Replicate weights are based on the same ideas as the jackknife, and has recently come into use in US government surveys, where replicate weights are provided instead of information on PSUs. In such cases, replicate weights may be used to disguise and/or prevent identification of individuals within PSUs to preserve privacy.

### 3.4.3.2.1 Handling of missing data

Missing values for the variables ALCEU, COCEU, ..., NUMTE in the file replicwts.LSF are coded −9.0. For these variables, 20.8% of the possible values are not observed. Use of listwise deletion would result in retaining only 97 of the selected 5005 cases. Therefore, we use the full information maximum likelihood (FIML) procedure as described in Section 3.6.

To define -9.0 as the global missing value code, select the **Data, Define Variables** option from the main menu. Select the variable ALCEU (or any other variable in the list) and click the **Missing Values** button to activate the **Missing Values for** dialog box. Type -9.0 in the **Global missing value** text box and select **pairwise** as the deletion method. Click **OK** when done.
3.4.3.2.2 **Handling of zero weights**

Descriptive statistics of the weight variables A2TWA0 to A2TWA78 revealed that these variables contained zero as possible values. If a zero value is encountered in any row of the data set, it was replaced by the average value of all the non-zero weights in the corresponding row.
We found that multiple imputation failed in estimating weights if 0 was regarded as a missing value code. The reason for this is that all the weight variables are highly correlated and therefore the covariance matrix of the weight variables is essentially singular.

To define CENREG and FACTYPE as stratification and cluster variables, select the **Data, Survey Design** option from the main menu and add the variables as shown in the dialog box below. For our first analysis, select A2TWA0 as the **Design weight** variable.

### 3.4.3.3 The model

The path diagram representation of the model fitted to the data is shown below. It is assumed that each of the X-variables AGE, GENDER, RACE_D, DEPR, EDU and JAILR is a perfect indicator of the corresponding latent variable, these being age, gender, race_d, depr, educ and jailr. This implies that the error variances of the X-variables are zero, and that the path coefficients age → AGE, gender → GENDER, race_d → RACE_D, depr → DEPR, educ → EDU and jailr → JAILR are all equal to one. In the LISREL terminology, age, gender, race_d, depr, edu and jail are exogenous (KSI) latent variables. In principle, several indicators of depression and education, if available, could be incorporated into this model.
In the Y part of the model, we include the dependent variables ALCEU, COCEU, MAREU and NUMTE. It is assumed that ALCEU, COCEU, and MAREU are indicators of the endogenous latent (ETA) variable subabuse while NUMTE is a perfect indicator of the ETA variable numte. The path subabuse → ALCEU is set equal to 1 to fix the scale of the endogenous latent variable subabuse.

Finally, we assume that numte can be predicted by age, ..., jailr, and in turn, subabuse is predicted by numte. In this context, the latent variable numte is a so-called mediating variable.

**Mathematical model**

The LISREL model consists of a measurement and structural part.

**Measurement model**

\[
\begin{pmatrix}
\text{y} \\
\text{x}
\end{pmatrix} = \begin{pmatrix}
\Lambda_y & 0 \\
0 & \Lambda_x
\end{pmatrix} \begin{pmatrix}
\eta \\
\xi
\end{pmatrix} + \begin{pmatrix}
\epsilon \\
\delta
\end{pmatrix}
\]

where \( x = (AGE, GENDER, RACE_D, DEPR, EDU, JAILR)' \), \( \Lambda_x \) is a 6×6 identity matrix and \( \xi = (age, gender, race_d, depr, edu, jailr)' \). Also, \( \text{Cov}(\delta) = 0 \) and \( \text{Cov}(\xi) = \Phi \).

Furthermore,
Finally, \( \text{Cov}(\varepsilon) \) is a diagonal matrix with diagonal elements \( \text{var}(ALCEU) \), \( \text{var}(COCEU) \), \( \text{var}(MAREU) \) and 0.

**Structural equation model**

The structural model can be written as

\[
\eta = \mathbf{B}\eta + \Gamma\xi + \varepsilon,
\]

where

\[
\mathbf{B} = \begin{bmatrix} 0 & \beta_{12} \\ 0 & 0 \end{bmatrix},
\]

and

\[
\Gamma = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & \gamma_{24} & \gamma_{25} & \gamma_{26} \end{bmatrix},
\]

that is,

\[
\text{subabuse} = \beta_{12} \times \text{numte} + t_i
\]
\[
\text{numte} = \gamma_{21} \times \text{age} + \gamma_{22} \times \text{gender} + \gamma_{23} \times \text{race} - d + \gamma_{24} \times \text{depr} + \gamma_{25} \times \text{educ} + \gamma_{26} \times \text{jailr}.
\]
Also

\[ \text{Cov}(\epsilon) = \Psi \]
\[ = \begin{bmatrix}
\psi_{11} & \psi_{21} \\
\psi_{21} & \psi_{22}
\end{bmatrix}. \]

The unknown model parameters are therefore \( \lambda_{y,21}, \lambda_{y,31}, \beta_{12}, \gamma_{21}, \gamma_{22}, \gamma_{23}, \gamma_{24}, \gamma_{25}, \gamma_{26}, \phi_{11}, \phi_{21}, \phi_{22}, \ldots, \phi_{66}, \psi_{11}, \psi_{21}, \psi_{22}, \text{var}(\epsilon_1), \text{var}(\epsilon_2), \text{and} \ \text{var}(\epsilon_3). \) In subsequent output, these 36 parameters will be denoted by the symbols LY21, LY31, BETA12, GAMMA21, GAMMA22, \ldots, GAMMA26, PHI11, PHI21, \ldots, PHI66, PSI11, PSI21, PSI22, TE11, TE22 and TE33.

### 3.4.3.4 Setting up the analysis using SIMPLIS syntax

It is relatively easy to specify the model described above with SIMPLIS syntax. We start by indicating that the raw data is to read from the file `replicwts.LSF`. Note that this is followed by a `REWIND` command which allows LISREL to repeatedly read the raw data from the same file.

The program commands `AGE = 1.0*age, ..., JAILR=1.0*jailr` specify that each of the X-variables are assumed to be exactly equal to the corresponding latent variable. Note that the part `1.0*latent variable` indicates that the path coefficient is fixed at the value of 1.0. In contrast, a command such as `COCEU = (0.5)*subabuse` indicates that 0.5 is a starting value (preliminary estimate) of \( \lambda_{y,21}. \) Since we assume that the X-variables measure the corresponding KSI latent variables without error, we set the error variances of NUMTE to JAILR to 0.
Next, we assume that subabuse is predicted by numte, and, in turn, numte is predicted by age, gender, race_d, depr, educ and jailr. This part of the syntax is shown below.

![Replicate weights. Missing Values in dataset Raw Data from file replicwts.lsf REWIND Latent Variables subabuse numte age gender race_d depr educ jailr Relationships: ALCEU = 1.0*subabuse COCEU = subabuse MAREU = subabuse NUMTE = 1.0*numte AGE = 1.0*age GENDER = 1.0*gender RACE_D = 1.0*race_d DEPR = 1.0*depr EDU = 1.0*educ JAILR = 1.0*jailr Set the Error Variance of NUMTE to 0.0 Set the Error Variance of AGE to 0.0 Set the Error Variance of GENDER to 0.0 Set the Error Variance of RACE_D to 0.0 Set the Error Variance of DEPR to 0.0 Set the Error Variance of EDU to 0.0 Set the Error Variance of JAILR to 0.0](replic1.spl)

To allow for the estimation of the covariance of the errors between the ETA latent variables numte and subabuse, we SET the error covariance between these variables free.

![replic1.spl](replic1.spl)

Finally, we use the LISREL OUTPUT: command to specify the number of repetitions (RP = 1), number of decimals (ND = 4), admissibility check off (AD = OFF), and save the standard error estimates (SV = replic1.std) and parameter estimates (PV = replic1.par) to text files replic1.std and replic1.par respectively.
3.4.3.5 Discussion of results

The estimated path coefficients for weight = A2TWA0 are shown in the path diagram given below. Although not presented here, all coefficients are statistically significant. The $\chi^2$-statistic for goodness of fit is 113.20, degrees of freedom is 19, and the RMSEA-value is 0.031.

Contents of the files replic1.par and replic1.std are given below. The estimates are preceded by three numbers $N_1$, $N_2$, and $N_3$, where $N_1 =$ repetition number, while $N_2$ and $N_3$ are zero if convergence has been attained. The parameter estimate and standard error values are given in scientific notation. For example,

$$0.165949D + 02 = 0.165942 \times 10^2 = 16.5942$$

$$0.234136D + 0 = 0.23413 \times 10^0 = 0.23413$$

In general, $D + k$ implies that the decimal point should be moved $k$ positions to the right, whereas $D - k$ implies the insertion of $k$ zero values just after the decimal point. For example $0.868283D - 03 = 0.000868283$. 
3.4.3.6 Use of replicate weights

We now demonstrate the use of replicate weights in LISREL, and start by making the LSF window the active window in order to invoke the LSF menu bar. From the Data menu, select Survey Design and remove the stratum variable CENREG and the cluster variable FACTYPE by clicking on the corresponding Remove buttons. Select the File, Save option to save this change.
Next, copy replic1.spl to replic2.spl and edit this SIMPLIS syntax file by changing $RP = 1$ to $RP = 79$ (shown in bold typeface below) and use the filenames replic2.std and replic2.par for saving the standard errors and parameter estimates respectively.

It is important to note that LISREL assumes that there are a total of 78 additional weight variables in the data set, starting with the **Design weight** variable name (in the present case A2TWA0) selected in the **Survey Design** dialog box. Additionally, it is assumed that the weight columns follow one another. Note that we are saving the estimated parameters and standard errors as Lisrel system data files (*.lsf).

### 3.4.3.7 Discussion of results

The results below show the parameter estimates and parameter standard error estimates for the model fitted with design weight A2TWA0, including stratification (CENREG) and clustering (FACTYPE) variables, and the corresponding average values using replicate weights without CENREG and FACTYPE.

Results of the parameters estimates for the first 10 repetitions are shown below.
Means of these estimates (and likewise for the standard error estimates) can be obtained by selecting **Statistics, Output Options...** from the main menu bar. This selection produces the **Output** dialog. Select **Covariances** and click **OK** to run PRELIS.

Portions of the output for the means of the parameter estimates and means of the standard error estimates are listed below.
In general, parameter estimates for the two estimation methods are quite close. Standard error estimates, on the other hand, tend to be larger for the replicate method. This is clearly a topic that requires further research.

3.5 Theory

3.5.1 Introduction

We assume that the population from which the sample data are obtained can be stratified into $H$ strata. Within each stratum $h$, $n_h$ clusters or primary sampling units (PSUs) are drawn and within the $h$-th stratum and $k$-th cluster, $n_{hk}$ ultimate sampling units (USUs) are drawn with design weights $w_{hk}$, where $l$ denotes the $l$-th USU within the $k$-th cluster, which in turn is nested within stratum $h$.

In the subsequent sections we discuss parameter estimation and the Taylor linearization method employed in LISREL to produce robust standard error estimates under single stage sampling.

3.5.2 Parameter estimation for continuous manifest variables

Assume that $y_{hd}$ is distributed as a $p \times 1$ multivariate normal random vector with mean $\mu$ and covariance matrix $\Sigma$. In structural equation modeling, it is hypothesized that $\Sigma = \Sigma(\gamma)$ and $\mu = \mu(\gamma)$, where $\gamma$ is a vector of $q$ unknown parameters to be estimated.
Example 1: CFA model with one factor

Consider the model $\Sigma(y) = \lambda \varphi \lambda' + D_o$. For this model, we have $\gamma = (\lambda_1, \lambda_2, ..., \lambda_p, \phi, \theta_1, \theta_2, ..., \theta_p)'$. To avoid any indeterminacy in estimating the unknown parameters, one would typically set the variance ($\phi$) of the latent factor equal to 1, or alternatively fix one of the factor loadings, for example $\lambda_1$, to 1. In the latter case, $\gamma = (\lambda_2, ..., \lambda_p, \psi, \theta_1, \theta_2, ..., \theta_p)'$.

Complete data

In the case of complete data, parameter estimation is relatively straightforward and can be summarized by the following two steps.

**Step 1**

Calculate the natural logarithm of the likelihood function, $\ln L$, where

$$\ln L = \sum_{k=1}^{H} \sum_{k=1}^{n_k} \sum_{l=1}^{m_k} w_{kkl} \ln f(y_{kkl} | \gamma)$$

(3.11)

where

$$f(y_{kkl} | \gamma) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp \left(-\frac{1}{2} tr\{\Sigma^{-1}(\gamma)G\}\right)$$

(3.12)

and

$$G = (y_{kkl} - \mu(\gamma))(y_{kkl} - \mu(\gamma))'$$

(3.13)

**Step 2**

Obtain an estimate $\hat{\gamma}$ of $\gamma$ by solving the set of simultaneous equations

$$\frac{\partial \ln L}{\partial \gamma} \bigg|_{\gamma=\hat{\gamma}} = 0$$

(3.14)

In general, no closed-form solution to the set of equations in (3.14) exists, and therefore parameter estimates are obtained iteratively using the Fisher scoring algorithm:
\[
\hat{\gamma}^{(t+1)} = \hat{\gamma}^{(t)} + \Gamma_n^{-1}(\hat{\gamma}^{(t)}) g(\hat{\gamma}^{(t)})
\]

(3.15)

where \(\hat{\gamma}^{(t)}\) denotes the parameter values at iteration \(t, t = 1, 2, \ldots\), and \(g(\cdot)\) denotes the gradient vector and where \(\Gamma_n(\cdot)\) denotes the Fisher information matrix for the elements of \(\gamma\).

In other words,

\[
g(\gamma) = \frac{\partial \ln L}{\partial \gamma}
\]

(3.16)

and

\[
\Gamma_n(\gamma) = -E \left[ \frac{\partial^2 \ln L}{\partial \gamma \partial \gamma'} \right]
\]

(3.17)

Iterations are continued until

\[
\left| \hat{\gamma}^{(t+1)} - \hat{\gamma}^{(t)} \right| < \epsilon \quad \forall i = 1, 2, \ldots, q
\]

where \(\epsilon\) is a small scalar value, e.g. \(10^{-6}\).

**Incomplete data**

Since the case of no missing values may be considered as a special case of a general framework that handles missing values, specific expressions for the gradient vector and information matrix will be given in this section.

Suppose that the data set contains \(n\) cases, where \(n = \sum \sum \sum 1\), then one could alternatively write (3.11) as

\[
\ln L = \sum_{\alpha=1}^{n} w_{\alpha} \ln f(y_{\alpha} | \gamma), \quad \alpha \text{ denotes the subscript } hkl.
\]

Suppose that the data can be split into two subsets of sizes \(n_1\) and \(n_2\), such that subset 1 contains no missing values and subset 2 contains no values for the \(p\)-th element of \(y_{\alpha}\), \(\alpha \in [\text{subset 2}]\). Let \(w_{ij}\) and \(y_{ij}\), \(j = 1, 2, \ldots, n_1\), denote the weights and observations in subset 1, while \(w_{2j}\) and \(y_{2j}\), \(j = 1, 2, \ldots, n_2\), denote the weights and observations in subset 2. In this case, we can write \(\ln L\) as

\[
\sum_{i=1}^{2} \sum_{j=1}^{n} w_{ij} \ln f(y_{ij} | \gamma)
\]

(3.18)
Suppose that, in general, there are \( \text{NPAT} \) patterns of missingness, \( i = 1, 2, \ldots, \text{NPAT} \) and that the number of observations within a pattern equals \( n_i, j = 1, 2, \ldots, n_i \). Let \( p_i \) indicate the number of variables with non-missing data.

**Example 2**

<table>
<thead>
<tr>
<th>Pattern</th>
<th>( y_1 )</th>
<th>( y_2 )</th>
<th>( y_3 )</th>
<th>( y_4 )</th>
<th>( p_i )</th>
<th>( w_{ij} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19.3</td>
<td>36.4</td>
<td>18.7</td>
<td>.</td>
<td>3</td>
<td>1.8</td>
</tr>
<tr>
<td>1</td>
<td>17.6</td>
<td>34.3</td>
<td>19.1</td>
<td>.</td>
<td>3</td>
<td>1.7</td>
</tr>
<tr>
<td>1</td>
<td>16.4</td>
<td>32.8</td>
<td>17.9</td>
<td>.</td>
<td>3</td>
<td>1.4</td>
</tr>
<tr>
<td>2</td>
<td>20.1</td>
<td>40.2</td>
<td>19.6</td>
<td>57.8</td>
<td>4</td>
<td>0.8</td>
</tr>
<tr>
<td>2</td>
<td>19.4</td>
<td>39.3</td>
<td>20.1</td>
<td>58.9</td>
<td>4</td>
<td>0.7</td>
</tr>
<tr>
<td>3</td>
<td>.</td>
<td>36.5</td>
<td>.</td>
<td>62.4</td>
<td>2</td>
<td>1.1</td>
</tr>
<tr>
<td>4</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>60.2</td>
<td>1</td>
<td>1.3</td>
</tr>
</tbody>
</table>

For the data shown above, \( \text{NPAT} = 4, n = 7, n_1 = 3, n_2 = 2, n_3 = 1 \) and \( n_4 = 1 \).

In general,

\[
\ln L = -\frac{1}{2} \sum_{i=1}^{\text{NPAT}} \sum_{j=1}^{n_i} w_{ij} \left\{ p_i \ln 2\pi + \ln |\Sigma_i| + tr\Sigma_i^{-1} (\bar{y}_i - \mu_i) (\bar{y}_i - \mu_i)' \right\}
\]

(3.19)

Let

\[
w_{i'} = \sum_{j=1}^{n_i} w_{ij}
\]

(3.20)

and

\[
\bar{y}_{w_i} = \frac{1}{w_i} \sum_{j=1}^{n_i} w_{ij} \bar{y}_{ij}
\]

(3.21)

**Example 3**

For the data set given in Example 1, \( w_3 = 1.8 + 1.7 + 1.4 = 4.9 \), and
Ignoring the weights, $\bar{y}_i = (17.77 \ 34.50 \ 18.57)$.

Use of (3.20) and (3.21) leads to the following expression for (3.19):

\[
\ln L = -\frac{1}{2} \sum_{i=1}^{NPAT} \left\{ w_i \left( p_i \ln 2\pi + \ln |\Sigma_i| \right) + \nu \Sigma_i^{-1} \sum_{j=1}^{n} w_j \left( y_{ij} y_j^i - y_{ij} \mu_j^i - \mu_j y_{ij}^i + \mu_j \mu_j^i \right) \right\} \\
= -\frac{1}{2} \sum_{i=1}^{NPAT} \left\{ w_i \left( p_i \ln 2\pi + \ln |\Sigma_i| \right) + \nu \Sigma_i^{-1} \left\{ \sum_{j=1}^{n} w_j y_{ij} y_j^i - w_i \bar{y}_w \mu_j^i - \mu_j w_i \bar{y}_w y_j^i - w_i \bar{y}_w \bar{y}_j^i + w_i \mu_j \mu_j^i \right\} \right\}
\]

Therefore

\[
\ln L = -\frac{1}{2} \sum_{i=1}^{NPAT} \left\{ w_i \left( p_i \ln 2\pi + \ln |\Sigma_i| \right) + \nu \Sigma_i^{-1} \left( G_w + S_w \right) \right\}
\]

(3.22)

where

\[
G_w = (\bar{y}_w - \mu_i)(\bar{y}_w - \mu_i)^T
\]

(3.23)

and

\[
S_w = \frac{1}{w_i} \sum_{j=1}^{n} w_j (y_{ij} - \bar{y}_w)(y_{ij} - \bar{y}_w)^T
\]

(3.24)

or equivalently,

\[
w_i S_w = \sum_{j=1}^{n} w_j y_{ij} y_j^i - w_i \bar{y}_w \bar{y}_j^i
\]

Gradient vector

\[
g(\gamma_r) = \sum_{i=1}^{NPAT} \sum_{j=1}^{n} w_j y_{ij} \left\{ (y_{ij} - \mu_i)^T \Sigma_i^{-1} \frac{\partial \mu_i}{\partial \gamma_r} + \frac{1}{2} \nu \Sigma_i^{-1} \left[ (y_{ij} - \mu_i)(y_{ij} - \mu_i)^T - \Sigma_i \right] \Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \gamma_r} \right\}
\]

(3.25)
Use of (3.20), (3.21), (3.22) and (3.23) gives

\[
g(\gamma_r) = \sum_{i=1}^{NPAT} \left\{ \sum_{j=1}^{n_i} \left[ (y_{ij} - w_{ij} \mu_i)^T \Sigma_i^{-1} \frac{\partial \mu_i}{\partial \gamma_r} + \frac{1}{2} tr \Sigma_i^{-1} \left( w_i S_{w_i} + w_i G_{w_i} - w_i \Sigma_i \right) \Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \gamma_r} \right] \right\}
\]

\[
= \sum_{i=1}^{NPAT} w_i \left\{ (y_{ij} - \mu_i)^T \Sigma_i^{-1} \frac{\partial \mu_i}{\partial \gamma_r} + \frac{1}{2} tr \Sigma_i^{-1} \left( S_{w_i} + G_{w_i} - \Sigma_i \right) \Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \gamma_r} \right\}
\]

(3.26)

Information matrix

\[
I_n(\gamma_r) = -E \left( \frac{\partial^2 \ln L}{\partial \gamma_r \partial \gamma_s} \right)
\]

\[
= \sum_{i=1}^{NPAT} \sum_{j=1}^{n_i} w_i \left\{ \frac{\partial \mu_i}{\partial \gamma_r} \Sigma_i^{-1} \frac{\partial \mu_i}{\partial \gamma_s} + \frac{1}{2} tr \Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \gamma_r} \Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \gamma_s} \right\}
\]

(3.27)

Let

\[
r_y = \Sigma_i^{-1} (y_{ij} - \mu_j)
\]

(3.28)

and

\[
P_y = r_y r_y' - \Sigma_i
\]

(3.29)

From (3.26) it follows that

\[
[g(\gamma)]_r = \sum_{i=1}^{NPAT} \sum_{j=1}^{n_i} [g_{ij}]_r
\]

(3.30)

where

\[
[g_{ij}]_r = w_i \left\{ r_{ij} \frac{\partial \mu_i}{\partial \gamma_r} + \frac{1}{2} tr \left\{ P_{ij} \frac{\partial \Sigma_i}{\partial \gamma_r} \right\} \right\}, \quad i = 1, 2, ..., NPAT; \quad j = 1, 2, ..., n_i
\]

(3.31)
Approximate covariance matrix of estimators

An approximate covariance matrix of the parameter estimators is derived as follows. From (3.14), (3.16) and (3.30) it follows that \( \hat{\gamma} \) is the solution to the set of equations

\[
\hat{w}(\hat{\gamma}) = \sum_{i=1}^{NPAT} \sum_{j=1}^{n_i} g_{ij}(\hat{\gamma}) = 0
\]  
(3.32)

Note that each \( g_{ij} \) is associated with a specific case \( hkl \), where \( h \) denotes strata, \( k \) clusters and \( l \) ultimate sample units. Using a first-order Taylor expansion of \( \hat{w}(\hat{\gamma}) \) at \( \hat{\gamma} = \gamma \), it follows that

\[
0 = \hat{w}(\hat{\gamma}) \approx \hat{w}(\gamma) + \frac{\partial \hat{w}(\gamma)}{\partial \gamma'} (\hat{\gamma} - \gamma)
\]  
(3.33)

Taking variances on both sides, it further follows that

\[
Cov(\hat{w}(\hat{\gamma})) \approx \frac{\partial \hat{w}(\gamma)}{\partial \gamma'} Cov(\hat{\gamma}) \left( \frac{\partial \hat{w}(\gamma)}{\partial \gamma'} \right)'.
\]  
(3.34)

Thus, provided that (cf. (3.32)) \( \frac{\partial^2 \ln L}{\partial \gamma \partial \gamma'} = \frac{\partial}{\partial \gamma} \left[ \frac{\partial g(\gamma)}{\partial \gamma'} \right] \) is a non-singular matrix,

\[
Cov(\hat{\gamma}) \approx \left[ \frac{\partial^2 \ln L}{\partial \gamma \partial \gamma'} \right]^{-1} Cov(\hat{w}(\hat{\gamma})) \left[ \frac{\partial^2 \ln L}{\partial \gamma \partial \gamma'} \right]^{-1},
\]

where \( E \left[ \frac{\partial^2 \ln L}{\partial \gamma \partial \gamma'} \right] = -I_n(\gamma) \).

Therefore, an approximate expression for the asymptotic covariance matrix of \( \hat{\gamma} \) is given by

\[
Cov(\hat{\gamma}) \approx I_n^{-1}(\gamma) G I_n^{-1}(\gamma)
\]  
(3.35)

where \( G = Cov(\hat{w}(\hat{\gamma})) \).
Using results derived by Binder (1983) and Fuller (1975), it follows that

\[
G = \frac{n-1}{n-q} \sum_{h=1}^{H} \frac{n_h (1 - f_h)}{n_h - 1} \sum_{i=1}^{n_h} \left( t_{h_i} - \bar{t}_{h_i} \right) \left( t_{h_i} - \bar{t}_{h_i} \right) ^{\prime},
\]

(3.36)

where

- \( n_h = \sum_{j=1}^{m_{hij}} \) the number of cases with identical response patterns within stratum \( h \), cluster \( i \), and USU \( j \). If \( f_{hij} = 1 \) for all \( h \), then \( m_{hij} = 1 \) for all \( h, i \) and \( j \).
- \( f_h = \frac{n_h}{N_h} \), the sampling rate for stratum \( h \).
- \( t_{hij} = g_{hij}(\hat{\gamma}) \), where \( g_{hij}(\hat{\gamma}) \) (cf. (3.31)) is the \( h \)-th contribution to the gradient vector \( g(\gamma) \) as defined by (3.30).
- \( t_{h_i} = \sum_{j=1}^{m_{hij}} t_{hij} \)
- \( \bar{t}_{h_i} = \frac{1}{n_h} \sum_{i=1}^{n_h} t_{h_i} \).

In practice, we assume a zero contribution to \( G \) for strata that contain a single PSU (cluster). Additionally, if the data do not contain a stratification variable, the PSUs are assumed to be the strata, and the observations (ultimate sampling units) within each PSU, clusters. Likewise, if there is no variable to define clusters, the observations within each stratum are treated as being the primary sampling units.

**Adjustment to the chi-square goodness of fit statistics**

Simulation studies indicated that the \( \chi^2_{LR} \)-statistic based on the log-likelihood (cf. (3.22)) in general yields too high a rejection rate. Let (cf. (3.35))

\[
d = tr \left( I_n(\hat{\gamma}) Cov(\hat{\gamma}) \right),
\]

(3.37)

Where \( I_n(\gamma) \) denotes the information matrix defined by (3.27).
A correction to the $\chi^2$-statistic for testing model fit is given by

$$
\chi^2_{\text{robust}} = c \times \chi^2_{\text{LR}},
$$

where

$$
c = \frac{q}{d},
$$

and where $q$ denotes the number of parameters to be estimated.

### 3.6 References


4 Structural equation models for a mixture of ordinal and continuous outcomes

4.1 Introduction

An important feature of software for the analysis of structural equation models (SEMs) is their facility to deal with a wide class of models for the analysis of latent variables (LVs). In the social sciences, and increasingly in biomedical and public health research, LV models have become an indispensable statistical tool.

There are basically three major reasons for the utility of LV models. First, this kind of model can summarize information contained in many response variables by a few latent variables. Second, when properly specified, a LV model can minimize the biasing effects of errors of measurement in estimating treatment effects. Third, LV models investigate effects between primary conceptual variables, rather than between any particular set of ordinary response variables. This means that a LV model is often viewed as more appropriate theoretically than is a simpler model with response variables only.

This guide deals with the use of design weights to fit SEM models to a mixture of continuous and ordinal manifest variables with or without missing values with optional specification of stratum and/or cluster variables. It also deals with the issue of robust standard error estimation and the adjustment of the chi-square goodness of fit statistic.

In the subsequent sections we briefly discuss parameter estimation (see du Toit, (2013)) and the Taylor linearization method employed in LISREL to produce robust standard error estimates under single stage sampling.

Two examples are given in Section 4.7. Additional examples are provided in the \orfimlex and \ls9ex subfolders of the LISREL Examples folder.

Research was supported by SBIR grant 5R44AA014999-03 from NIAAA to Scientific Software International.

4.2 A re-parameterization of the general LISREL model

In this section, a re-parameterization of the general LISREL model (Jöreskog, K.G & Sörbom, D., 2004) is presented. The relationships between the parameters of the model and the LISREL parameter matrices are also described.

Let $y_i$ be a $(p+q)$ vector of latent variables and consider the general model

$$y_i = u + \Lambda (\alpha_i + u_i) + \varepsilon_i,$$

where
\( \mathbf{v} = \begin{pmatrix} \tau_y : (px1) \\ \tau_x : (qx1) \end{pmatrix} \), \hspace{1cm} (4.2)

\[
\Lambda = \begin{pmatrix} \Lambda_y (I - B)^{-1} & \Lambda_y (I - B)^{-1} \Gamma \\ 0 & \Lambda_x \end{pmatrix}
\]

\( \alpha = \begin{pmatrix} \alpha_y \\ \kappa \end{pmatrix} \), \hspace{1cm} (4.4)

\[
\mathbf{u}_i = \begin{pmatrix} \mathbf{w}_i \\ \mathbf{v}_i \end{pmatrix}
\]

\[
\text{Cov}(\mathbf{u}_i) = \Phi = \begin{pmatrix} \Psi_y & 0 \\ 0 & \Phi \end{pmatrix}
\]

\( \mathbf{e}_j = \begin{pmatrix} \mathbf{e}_y \\ \mathbf{e}_\delta \end{pmatrix} \), \hspace{1cm} (4.7)

\[
\text{Cov}(\mathbf{e}_j) = \Theta = \begin{pmatrix} \Theta_e & \Theta_{ed} \\ \Theta_{de} & \Theta_d \end{pmatrix}
\]

From the definitions above it follows that:

- \( \Lambda_y \) of the general LISREL model is part of the first \( p \times m \) submatrix of \( \Lambda \)
- \( \Lambda_x \) of the general LISREL model is the final \( q \times n \) submatrix of \( \Lambda \)
- \( \mathbf{B} \) of the general LISREL model is an \( m \times m \) matrix with zero diagonal elements and is part of the first \( p \times m \) and second \( p \times n \) submatrices of \( \Lambda \)
- \( \Gamma \) of the general LISREL model is an \( m \times n \) submatrix of \( \Lambda \)
- \( \tau_y \) of the general LISREL model consists of the first \( p \) elements of \( \mathbf{v} \)
• $\boldsymbol{\tau}_s$ of the general LISREL model consists of the final $q$ elements of $\upsilon$

• $\boldsymbol{a}_y$ of the general LISREL model consists of the first $m$ elements of $\alpha$

• $\boldsymbol{k}$ of the general LISREL model consists of the last $n$ elements of $\alpha$

• $\Theta_\varepsilon$ of the general LISREL model is the first $p \times p$ submatrix of $\Theta$

• $\Theta_\delta$ of the general LISREL model is the final $q \times q$ submatrix of $\Theta$

• $\Theta_{\varepsilon\delta} = \Theta'_{\delta\varepsilon}$ of the general LISREL model is the second $p \times q$ submatrix of $\Theta$

• $\Phi$ of the general LISREL model is the final $n \times n$ submatrix of $\Psi$

• $\Psi_y$ of the general LISREL model is the first $m \times m$ submatrix of $\Psi$

4.3 Mixture of ordinal and continuous variables

Let

$$
\mathbf{y}_i = \begin{pmatrix}
\mathbf{y}_{0i} \mid p_i \times 1 \\
\mathbf{y}_{Ni} \mid q_i \times 1
\end{pmatrix}, \quad i = 1, 2, \ldots, N
$$

(4.9)

where $N$ denotes the number of cases and where the $(p_i + q_i) \times 1$ vector $\mathbf{y}_i$ of manifest variables are partitioned into a $p_i \times 1$ vector $\mathbf{y}_{0i}$ of ordinal and a $(q_i \times 1)$ vector $\mathbf{y}_{Ni}$ of continuous manifest variables.

It is further assumed that the SEM model has $m$ latent variables $\mathbf{u}_i$, where $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$ are i.i.d. $N(\mathbf{0}, \Psi)$.

The likelihood function for case $i$ is evaluated as

$$
f(\mathbf{y}_i) = \int f(\mathbf{y}_{0i}, \mathbf{y}_{Ni}, \mathbf{u}_i) d\mathbf{u}_i \\
= \int f(\mathbf{y}_{0i}, \mathbf{y}_{Ni} \mid \mathbf{u}_i) g(\mathbf{u}_i) d\mathbf{u}_i.
$$

(4.10)

Under the assumption of conditional independence, it follows that
\[ f(y_i) = \prod_{j=1}^{p_i} f(y_{ij} | u_j) \cdot \prod_{j=1}^{q_i} f(y_{xij} | u_j) g(u_j) \, du_j. \]  

(4.11)

Hence

\[ f(y_i) = \int \exp \left\{ \sum_{j=1}^{p_i} \ln f(y_{ij} | u_j) + \sum_{j=1}^{q_i} \ln f(y_{xij} | u_j) + \ln g(u_j) \right\} \, du_j. \]  

(4.12)

In general, a closed-form solution to this integral does not exist. To evaluate integrals of the type described above, we use a direct implementation of Gauss-Hermite quadrature.

With this rule, an integral of the form

\[ I(t) = \int f(t) \exp \left[ -t^2 \right] \, dt \]

is approximated by the sum

\[ I(t) \approx \sum_{u=1}^{Q} w_u f(z_u), \]

where \( w_u \) and \( z_u \) are weights and nodes of the Hermite polynomial of degree \( Q \).

Adaptive quadrature generally requires fewer points and weights to yield estimates of the model parameters and standard errors that are as accurate as would be obtained with more points and weights in non-adaptive quadrature. The reason for that is that the adaptive quadrature procedure uses the empirical Bayes means and covariances, updated at each iteration to essentially shift and scale the quadrature locations of each case (subject) in order to place them under the peak of the corresponding integral.

\[ f(y_{ij} | u_j): \text{Ordinal variables} \]

Suppose that \( y_{ij} \) has \( C \) categories, then

\[ P(y_{ij} = c) = P(y_{ij} \leq c) - P(y_{ij} \leq c - 1), \quad c = 1, 2, \ldots, C - 1 \]

where \( P(y_{ij} = 0) = 0 \) and \( P(y_{ij} = C) = 1 - P(y_{ij} \leq c - 1). \)

In LISREL the logit, probit, log-log and cumulative log-log functions are available.
In the case of ordinal variables it is assumed that the corresponding subsets of \( \upsilon \) and \( \Theta \) are set to zero means and unit variances. The parameter \( \tau_{ic} \) is the so-called threshold parameter.

\[ f(y_{Nij} \mid u_j): \text{Normal variables} \]

Suppose that \( y_{Ni} \) has typical element \( y_{Nij} \), then

\[ f(y_{Nij} \mid u_j) = \left(2\pi\theta_{jj}\right)^{-1/2}\exp\left(-\frac{1}{2\theta_{jj}}(y_{Nij} - \mu_{ij})^2\right) \]

where

\[ \mu_{ij} = v_j + \lambda_i (a + u_i) \]

The parameter \( v_j \) is the so-called intercept parameter and \( \theta_{jj} \) the residual variance.

See du Toit (2013) for the treatment of the special case where the latent variables associated with the continuous variables and those associated with the ordinal variables are mutually exclusive.

### 4.4 Parameter estimation

Parameter estimation is relatively straightforward and can be summarized by the following two steps.

Step 1:

Calculate the natural logarithm of the likelihood function, \( \ln L \), where

\[ \ln L = \sum_{k=1}^{H} \sum_{k=1}^{n_k} \sum_{l=1}^{n_l} w_{kl} \ln f(y_{kl} \mid \gamma). \tag{4.13} \]

Step 2:

Obtain an estimate \( \hat{\gamma} \) of \( \gamma \) by solving the set of simultaneous equations
In general, no closed-form solution to the set of equations (18) exists, and therefore parameter estimates are obtained interactively using the Fisher scoring algorithm:

\[ \hat{\gamma}^{(t+1)} = \hat{\gamma}^{(t)} + I_n^{-1}(\hat{\gamma}^{(t)}) g(\hat{\gamma}^{(t)}) \]  

(4.15)

where \( \hat{\gamma}^{(t)} \) denotes the parameter values at iteration \( t \), \( t = 1, 2, \ldots \), and \( g(\cdot) \) denotes the gradient vector and where \( I_n(\cdot) \) denotes the information matrix.

Suppose that the elements of \( \gamma : (k \times 1) \) are functions of the elements of \( \gamma^*: (m \times 1) \), where

\[ \gamma^* = vec \left[ \Lambda_y (I - B)^{-1} \Lambda_y (I - B)^{-1} \Gamma \right]. \]

Suppose further that it is possible to evaluate \( \frac{\partial \ln L}{\partial \gamma^*} = g(\gamma^*) \) and \( \frac{\partial^2 \ln L}{\partial \gamma \partial \gamma} = H(\gamma^*). \)

Proposition 1

\[ g(\gamma) = C'g(\gamma^*) \]

where

\[ C : m \times k \text{ has typical column } C_i = \frac{\partial \gamma^*}{\partial \gamma_i}, i = 1, 2, \ldots, k \]

Proposition 2

Let

\[ I_n(\gamma) = -E[H(\gamma)], \]

where
\[
[H(\gamma)]_{ij} = \frac{\partial^2 \ln L}{\partial \gamma_i \partial \gamma_j}.
\]

Then

\[
\text{vecs}[H(\gamma)] = S g(\gamma') + \text{vecs} \left[ C'H(\gamma')C \right],
\]

where \( S: (k^* \times m) \) has typical row \( \frac{\partial^2 \gamma^*}{\partial \gamma_i \partial \gamma_j}, i = 1, 2, \ldots, k; j = 1, 2, \ldots, i \) and \( k^* = k(k+1)/2 \).

Proofs of these propositions are provided in du Toit (2013).

Iterations are continued until \( \gamma^{(i+1)} - \gamma^{(i)} \prec \varepsilon \) \( \forall i = 1, 2, \ldots, n_{\text{free}} \) where \( \varepsilon \) is a small scalar value, e.g. \( 10^{-6} \).

### 4.5 Approximate covariance matrix of estimators

There are a number of basic methods available (see, e.g., Rao, 1975) for estimation of the sampling variances and covariances of the estimated parameters. The method used in LISREL is based on the linearization methods described by Wolter (1985) and Binder (1983).

An approximate expression for the asymptotic covariance matrix of \( \hat{\gamma} \) is given by

\[
\text{Cov}(\hat{\gamma}) \approx I_n^{-1}(\gamma)G I_n^{-1}(\gamma)
\]

Using results derived by Binder (1983) and Fuller (1975), it follows that, under single stage sampling with replacement (WR) or without replacement (WOR)

\[
G = \sum_{h=1}^{H} n_h (1-f_h) \sum_{i=1}^{n_h} (t_{hi} - \bar{t}_h)(t_{hi} - \bar{t}_h)
\]

(4.18)

where
\[ n_h = \sum_{j=1}^{n_{hj}} m_{hij}, \text{ with } m_{hij} \text{ the number of cases within stratum } h, \text{ cluster } i, \text{ and USU } j. \]

\[ f_h = \frac{n_h}{N}, \text{ the sampling rate for stratum } h. \]

\[ t_{hij} = g_{hij}(\hat{\gamma}), \text{ where } g_{hij}(\hat{\gamma}) \text{ is the } h_{ij}-\text{th contribution to the gradient vector } g(\gamma). \]

\[ t_{h_i} = \sum_{j=1}^{m_{hij}} t_{hij} \]

\[ \bar{t}_{h_i} = \frac{1}{n_h} \sum_{i=1}^{n_h} t_{h_i}. \]

In practice, we assume a zero contribution to \( G \) for strata that contain a single PSU (cluster). Additionally, if there is no variable to define clusters, the observations within each stratum are treated as being the primary sampling units.

### 4.6 Adjustment to the chi-square goodness of fit statistics

Simulation studies indicated that the \( \chi^2_{LR} \)-statistic based on the difference between two deviance statistics in general yields a too high rejection rate.

The chi-square adjustment described below is based on contributions by Satorra, A. and Bentler, P.M. (1994, 2010).

Let

\[ d_1 = \text{tr} \left( I_n \left( \hat{\gamma}_1 \right) \text{Cov} \left( \hat{\gamma}_1 \right) \right), \quad d_2 = \text{tr} \left( I_n \left( \hat{\gamma}_2 \right) \text{Cov} \left( \hat{\gamma}_2 \right) \right) \quad (4.19) \]

where \( I_n \left( \gamma_s \right), s = 1, 2 \) denotes the information matrix under \( H_1 \) and \( H_2 \).

A correction to the \( \chi^2 \)-statistic for testing the difference in two deviance statistics is given by

\[ \chi^2_{\text{robust}} = c \times \chi^2_{LR} \quad (4.20) \]
where \((n_{free_2} > n_{free_1})\)

\[
c = \frac{n_{free_2} - n_{free_1}}{\text{abs}(d_1 - d_2)},
\]

(4.21)

and where \(n_{free_1}\) and \(n_{free_2}\) respectively denote the total number of parameters to be estimated under the \(H_1\) and \(H_2\) models.

4.7 Examples

4.7.1 Confirmatory factor analysis model

The data set (\texttt{drug_select.lsf}) forms part of the data library of the Alcohol and Drug Services Study (ADSS). The ADSS is a national study of substance abuse treatment facilities and clients. Background data and data on the substance abuse of a sample of 1752 clients were obtained. The sample was stratified by census region and within each stratum a sample was obtained for each of three facility treatment types within each of the four census regions.

The following variables included in the LSF were selected from the survey data:

- \texttt{CENREG}: This variable indicates the census region and has four categories, these being "Northeast", "Midwest", "South", and "West" respectively.
- \texttt{FACTYPE}: The facility treatment type has four categories, too, representing facilities with "residential treatment", "outpatient methadone treatment", "outpatient non-methadone treatment", and "more than one type of treatment" respectively.
- \texttt{COCEU}: An indicator variable with value "1" if the respondent has ever used cocaine, and "0" otherwise.
- **MAREU**: An indicator variable with value "1" if the respondent has ever used marijuana, and "0" otherwise.
- **DEPR**: This indicator variable is coded "1" if the respondent is depressed, and "0" otherwise.
- **EDU**: A categorical variable representing the respondent's level of education at admission. It has 5 categories, these being (from 1 to 5) "less than 8 years", "8 – 11 years or less than High School graduate", "High School graduate / GED", "some college", and "college graduate / postgraduate".
- **JAILR**: This indicator variable indicates whether the respondent had a prison or jail record prior to admission.
- **NUMTE**: A count variable, indicating the total number of treatment episodes prior to admission.

From the main menu bar, select the **Data, Survey Design … option**.

Add the survey design variables to the appropriate text boxes and select the **File, Save** option to save these changes to the LSF file.
The SIMPLIS syntax file is shown below. The only difference between this and the usual SIMPLIS syntax is the addition of the paragraph

$ADAPQ(12) CLL GR(5)

See the SIMPLIS Syntax Guide for a description of the $ADAPQ(n) command and options.

A portion of the LISREL output is shown next.
Fit statistics and threshold values are shown below.

<table>
<thead>
<tr>
<th></th>
<th>DrugUse</th>
<th></th>
<th>Environm</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>COCEU</td>
<td>1.494</td>
<td>(0.543)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>2.753</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAREU</td>
<td>1.447</td>
<td>(0.150)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>0.872</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JAILR</td>
<td>0.754</td>
<td>(0.188)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>4.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEPR</td>
<td>-</td>
<td>0.341</td>
<td></td>
<td>0.102</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.351</td>
</tr>
<tr>
<td>EDU</td>
<td>-</td>
<td>-0.133</td>
<td></td>
<td>0.082</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-1.627</td>
</tr>
<tr>
<td>NUMTE</td>
<td>-</td>
<td>4.329</td>
<td></td>
<td>1.403</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.086</td>
</tr>
</tbody>
</table>
A path diagram representation of the model fitted to the data is shown below. From the path diagram display it can be seen, for example, that the correlation between the two latent variables equals 0.487.
4.7.2 Confirmatory factor analysis model with latent variable relationship and latent variable means

This example is based on six political efficacy measurements as described in Aish & Jöreskog (1990). The dataset EFFICACY.LSF (orfimlex folder) consists of 1719 cases obtained in a USA sample where the number -999999.0 denotes a missing value.
The data are the responses to the following statements:

- **NOSAY**: "People like me have no say in what the government does."
- **VOTING**: "Voting is the only way that people like me can have any say about how the government runs things."
- **COMPLEX**: "Sometimes politics and government seem so complicated that a person like me cannot really understand what is going on."
- **NOCARE**: "I don't think that public officials care much about what people like me think."
- **TOUCH**: "Generally speaking, those who elect to Congress in Washington lose touch with the people pretty quickly."
- **INTEREST**: "Parties are only interested in people's votes but not in their opinions."

The ordered categories are:

- 1: agree strongly
- 2: agree
- 3: disagree
- 4: disagree strongly

It is hypothesized that the six variables are indicators of two factors, where Factor 1 represents Internal Efficacy measured by NOSAY, VOTING and COMPLEX. Factor 2 represents External Efficacy measured by NOCARE, TOUCH and INTEREST.

It is further hypothesized that Factor 2 depends on Factor 1. The LISREL syntax for fitting the model is shown next. Six quadrature points are selected for each of the two dimensions using the probit link function.
A portion of the output is shown next, being the parameter specifications where values greater than zero correspond to parameters that are estimated while zero values correspond to parameters that are fixed at zero or the values entered in the "VA" paragraphs.
The path diagram representation shown below provides the values of the estimated parameters for factor means.
The path diagram for the basic model showing the parameter estimates are given next.
4.8 References and further reading


