



## Simulation with PRELIS and LISREL

### Contents

1. Introduction .....	1
2. Bootstrapping.....	2
3. Monte Carlo experiments .....	5
3.1 Generating variables with a specified covariance matrix: Monte Carlo example 1 .....	5
3.2 Generating normal variables .....	6
3.3 Generating non-normal variables: Monte Carlo example 2 .....	9
3.4 Generating variables from a specified model.....	10
3.5 Monte Carlo example 3.....	11
3.6 Generating ordinal variables: Monte Carlo example 4 .....	12
4. References .....	17

### 1. Introduction

By simulation we mean here drawing random samples of data from some population, estimating various parameters from each sample for the purpose of studying the mean and variance and other characteristics of the distribution of these parameter estimates. There are two techniques for drawing the random samples: bootstrap and Monte Carlo. In bootstrapping the random samples are drawn from an original sample, which usually is a sample of empirical data but which can also be a set of artificial data. In Monte Carlo sampling, the samples are generated from randomly generated variables so no real data is involved. Various combinations of the two techniques are also possible.

In this document we illustrate how bootstrapping and Monte Carlo experimenting can be done with PRELIS and LISREL. We emphasize the interplay between the two programs. Typically, the random samples are generated by PRELIS. These samples are then analyzed by LISREL. The sampling distribution of the parameter estimates or other quantities produced by LISREL are then studied by PRELIS.

The quantities which distributions can be studied in this way are parameter estimates, goodness-of-fit statistics, fitted variances, covariances or correlations, factor score regressions, and estimated asymptotic covariance matrices of parameter estimators and these can be studied under any combination of *type of matrix analyzed* (MA = AM, CM, KM, MM, CM, PM, RM, TM) and *method of estimation* (ME = IV, TSLS, ULS, GLS, ML, WLS, DWLS). The vector of estimated free parameters can be saved by putting PV = *filename1* and the vector of corresponding standard errors can be saved by putting SV = *filename2* on the OU command. The *t*-values may be saved by putting TV = *filename3* on the OU command.

We illustrate these procedures on several examples in such a way that anybody interested can carry out these experiments directly and use any variation of them as required. The data file for the bootstrap example is **EFFICACY.RAW** and the input files **BS\_ST1.PRL**, **BS\_ST2.LIS** and **BS\_ST3.PRL**. For the Monte Carlo experiments the input files are denoted **SIMEXij,XXX**, where  $i$  is the example number,  $j$  is a sequence number within the example, and  $XXX$  is PRL for a PRELIS input file and LIS for a LISREL input file. To avoid excessive output files in simulations, put  $XO$  or  $XO = 0$  on the OU command. All these files can be found in the **PRELIS Examples** folder.

For both bootstrapping and Monte Carlo experimenting, we use the random number generator of Schrage (1979) to generate uniform random numbers in double precision. This requires a seed as a starting number for the generator. The seed may be specified as  $IX = i$  on the OU command of the PRELIS input file, where  $i$  is a random integer. If  $IX$  is not specified, its value will be set equal to the number of seconds passed since last midnight. In this case, this means that if two persons run the same input file or if one person runs the same input file at different occasions, the two runs will not produce the same result, although the two results represent realizations of the same random phenomenon. To reproduce exactly the same result as reported here, one must use the same seed as we have used.

## 2. Bootstrapping

The original data consists of  $N$  cases and we want to draw  $K$  samples of size  $n$ . The drawing is done with replacement. The number  $n$  may be smaller than, equal to, or larger than  $N$ . For each of the  $K$  samples, some statistics are computed and saved in a file. We assume here that the data matrix of order  $N \times k$  does not contain missing values. (If it does, just specify the missing values in the input file in Step 1; all bootstrap samples will be drawn from the listwise sample remaining after deletion of all cases with missing values.)

Suppose we want to study the distribution of the factor loadings (some or all of them) in LISREL when a specified model is estimated by ML on the basis of Kendall's tau-c correlations. There is no statistical theory that justifies using ML to fit a model to a matrix of tau-c correlations. In particular, one cannot rely on standard errors of parameter estimates. However, with the bootstrap technique one can estimate valid standard errors in this case.

To illustrate, we draw 100 bootstrap samples of size 148 from the raw data in the file **efficacy.raw** containing 297 cases on six variables. The data and syntax for these examples can be found in the **PRELIS examples\Imputation and Bootstrap** folder.

### Step 1

Run PRELIS with the following input file (**BS\_ST1.PRL**):

```
!Generating 100 TM matrices (Kendall's tau-c correlations)
!by bootstrapping from EFFICACY.RAW
DA NI=6
LA=EFFICACY.LAB
RA=EFFICACY.RAW
OU SP MA=TM BS=100 SF=50 BM=EFFICACY.TMB IX=1234567
```

Here,  $BS = 100$  is the number of bootstrap samples to be drawn and  $SF$  is the sample fraction as a percentage. With  $SF = 50$ , each bootstrap sample will be of size 148.

After this run, the file **EFFICACY.TMB** contains 100 matrices of tau-c correlations. The correlation matrices for the first two bootstrap samples are:

0.10000D+01	0.19490D+00	0.10000D+01	0.18604D+00	0.15559D+00	0.10000D+01
0.29574D+00	0.17513D+00	0.25217D+00	0.10000D+01	0.19791D+00	0.12929D+00
0.20823D+00	0.39068D+00	0.10000D+01	0.23722D+00	0.14270D+00	0.23811D+00
0.41618D+00	0.41942D+00	0.10000D+01			
0.10000D+01	0.18801D+00	0.10000D+01	0.14401D+00	0.13706D+00	0.10000D+01
0.33937D+00	0.13769D+00	0.24050D+00	0.10000D+01	0.24465D+00	0.13179D+00
0.17843D+00	0.38008D+00	0.10000D+01	0.29275D+00	0.12562D+00	0.20397D+00
0.39593D+00	0.42876D+00	0.10000D+01			

## Step 2

Each of the 100 correlation matrices generated in Step 1 will be analyzed by LISREL by fitting a specific two-factor model. The matrix of factor loadings  $\Lambda_x$  (LX) estimated in each sample will be saved in a file **EFFICACY.LXB**. The LISREL input file (**BS\_ST2.LIS**) has the following form:

```
!Estimating 100 LX matrices (factor loadings) from EFFICACY.TMB
DA NI=6 NO=148 RP=100
LA=EFFICACY.LAB REWIND
KM=EFFICACY.TMB
MO NX=6 NK=2
FR LX(1,1) LX(2,1) LX(3,1) LX(4,1) LX(4,2) LX(5,2) LX(6,2)
OU LX=EFFICACY.LXB XM
```

The RP = 100 means *repeat 100 times* and has the same effect as stacking the 100 input files after each other, *i.e.*, the input file above will be read 100 times. Note that the labels file (**EFFICACY.LAB**) is rewound after each sample has been analyzed. The XM option on the OU command tells LISREL not to compute modification indices, as this saves some computer time.

This run will produce a huge output file containing the printed output for each of the 100 analyses. If these results are of no interest, one can avoid this by putting XO or XO = 0 on the OU command. XO generates output for the first repetition; XO = 0 generates no output at all.

This step produces the file **EFFICACY.LXB** containing 100 matrixes of factor loadings. The set of factor loadings for the first two bootstrap samples are:

0.48969D+00	0.00000D+00	0.47880D+00	0.00000D+00	0.26696D+00	0.00000D+00
0.31653D+00	0.52600D+00	0.00000D+00	0.63859D+00	0.00000D+00	0.70994D+00
0.45676D+00	0.00000D+00	0.15253D+00	0.00000D+00	0.41521D+00	0.00000D+00
0.14165D+00	0.60094D+00	0.00000D+00	0.61809D+00	0.00000D+00	0.65081D+00

## Step 3

The empirical distribution of the 100 sets of factor loadings in **EFFICACY.LXB** generated in Step 2 can be studied by running PRELIS, treating the factor loadings as continuous variables. PRELIS gives the mean, variance, skewness, kurtosis, maximum, and minimum for each loading. Univariate and multivariate tests of normality are also given.

The PRELIS input file (**BS\_ST3.PRL**) for step 3 is:

```
!Analyzing 100 LX matrices (factor loadings) in EFFICACY.LXB
DA NI=12
LA
'LX(1,1)' 'LX(1,2)' 'LX(2,1)' 'LX(2,2)' 'LX(3,1)' 'LX(3,2)' 'LX(4,1)'
```

```
'LX(4,2)' 'LX(5,1)' 'LX(5,2)' 'LX(6,1)' 'LX(6,2)'
RA=EFFICACY.LXB
CO ALL
SD 'LX(1,2)' 'LX(2,2)' 'LX(3,2)' 'LX(5,1)' 'LX(6,1)'
OU MA=KM WP
```

Since the fixed zero factor loadings are of no interest, we use an SD command to eliminate these. The results are as follows:

#### Univariate Summary Statistics for Continuous Variables

Variable	Mean	St. Dev.	Skewness	Kurtosis	Minimum	Freq.	Maximum	Freq.
LX(1,1)	0.527	0.035	-0.107	-0.523	0.440	1	0.604	1
LX(2,1)	0.349	0.038	0.154	-0.043	0.260	1	0.461	1
LX(3,1)	0.413	0.032	-0.125	-0.786	0.343	1	0.481	1
LX(4,1)	0.313	0.070	0.051	0.559	0.120	1	0.496	1
LX(4,2)	0.383	0.070	-0.171	-0.171	0.208	1	0.538	1
LX(5,2)	0.604	0.025	-0.219	-0.151	0.550	1	0.664	1
LX(6,2)	0.665	0.024	-0.052	-0.466	0.605	1	0.717	1

#### Test of Univariate Normality for Continuous Variables

Variable	Skewness		Kurtosis		Skewness and Kurtosis	
	Z-Score	P-Value	Z-Score	P-Value	Chi-Square	P-Value
LX(1,1)	-0.458	0.647	-1.296	0.195	1.890	0.389
LX(2,1)	0.657	0.511	0.091	0.928	0.440	0.803
LX(3,1)	-0.532	0.595	-2.450	0.014	6.287	0.043
LX(4,1)	0.218	0.827	1.198	0.231	1.483	0.476
LX(4,2)	-0.730	0.466	-0.218	0.827	0.580	0.748
LX(5,2)	-0.927	0.354	-0.167	0.867	0.888	0.641
LX(6,2)	-0.223	0.823	-1.093	0.275	1.243	0.537

Relative Multivariate Kurtosis = 0.954

#### Test of Multivariate Normality for Continuous Variables

Value	Skewness		Kurtosis			Skewness and Kurtosis	
	Z-Score	P-Value	Value	Z-Score	P-Value	Chi-Square	P-Value
4.582	-0.556	0.578	60.094	-0.713	0.476	0.817	0.665

#### Correlation Matrix

	LX(1,1)	LX(2,1)	LX(3,1)	LX(4,1)	LX(4,2)	LX(5,2)	LX(6,2)
LX(1,1)	1.000						
LX(2,1)	0.218	1.000					
LX(3,1)	-0.077	0.181	1.000				
LX(4,1)	0.062	-0.162	-0.036	1.000			
LX(4,2)	-0.015	0.174	0.048	-0.949	1.000		
LX(5,2)	0.111	0.139	-0.083	0.038	-0.071	1.000	
LX(6,2)	0.150	-0.123	-0.082	0.399	-0.358	-0.106	1.000

Means

LX(1,1)	LX(2,1)	LX(3,1)	LX(4,1)	LX(4,2)	LX(5,2)	LX(6,2)
-----	-----	-----	-----	-----	-----	-----
0.527	0.349	0.413	0.313	0.383	0.604	0.665

Standard Deviations

LX(1,1)	LX(2,1)	LX(3,1)	LX(4,1)	LX(4,2)	LX(5,2)	LX(6,2)
-----	-----	-----	-----	-----	-----	-----
0.035	0.038	0.032	0.070	0.070	0.025	0.024

The standard deviations are the standard errors of factor loadings that one should expect with this procedure for samples of size 148. It is seen that most loadings can be estimated with reasonable precision by the combination of a TM matrix and the ML method, except LX(4,1) and LX(4,2), which have much higher standard errors and much higher kurtoses than the other loadings. Normality is also rejected for these loadings. Note also that the correlations of the factor loadings are generally small except for LX(4,1) and LX(4,2) which are almost perfectly linearly related. It is clear that with this method and this small sample size, these two loadings are not separately estimable.

### 3. Monte Carlo experiments

Monte Carlo experiments can be done in a similar way as bootstrapping. The difference is that there is no raw data to start with. However, raw data can be generated directly with PRELIS, and covariance or correlation matrices can be computed directly *without saving or storing the raw data*.

The key elements here are the two random variables NRAND and URAND. NRAND generates a random normal variable with mean zero and variance one. URAND generates a uniform random variable over the interval (0,1). Each time NRAND and URAND is specified in the input file, a new random normal variable independent of previously generated variables, is generated. Although we use capitalized names for NRAND and URAND here, PRELIS recognizes any combination of upper case and lower case characters in these names.

In PRELIS, the random variables NRAND and URAND act as ordinary variables, except they are not read from a raw data file. Using the PRELIS command NE (for new variable), they can be combined with other variables which have been read from a file or have been generated previously. As the following examples demonstrate, there are almost unlimited possibilities of generating normal and non-normal variables with specified properties. Using the PRELIS RE (recode) command one can also generate discrete variables (ordinal or categorical). We give four typical examples. The last of these is given in full detail. All these examples may be found in the **PRELIS Examples\Simulated data examples** folder.

#### 3.1 Generating variables with a specified covariance matrix: Monte Carlo example 1

Suppose the population covariance matrix  $\Sigma$  is given, say, by

$$\Sigma = \begin{bmatrix} 1.000 & & & & & \\ 0.378 & 1.000 & & & & \\ 0.720 & 0.336 & 1.000 & & & \\ 0.324 & 0.420 & 0.288 & 1.000 & & \\ 0.270 & 0.350 & 0.240 & 0.300 & 1.000 & \\ 0.270 & 0.126 & 0.240 & 0.108 & 0.090 & 1.000 \end{bmatrix}$$

Although this is a correlation matrix, it is treated here as a covariance matrix.

The population covariance matrix  $\Sigma$  may or may not be specified such that a particular model fits this matrix exactly. In the first case, the *central* case, one can study how parameter estimates and other quantities behave when the model holds exactly in the population. In the second case, the *non-central* case, one can study the behavior when the model does not hold or holds only approximately in the population.

If  $\Sigma$  is positive definite, there exists a lower triangular matrix  $T$ , such that  $\Sigma = TT'$ . The matrix  $T$  can be computed by LISREL using the following input file (**montecarlo.lis**)

```
!Fitting TT' to Sigma
DA NI = 6 NO = 100000; CM SIGMA; MO NX = 6 NK = 6 PH = ID TD = ZE
PA LX
1 0 0 0 0 0
1 1 0 0 0 0
1 1 1 0 0 0
1 1 1 1 0 0
1 1 1 1 1 0
1 1 1 1 1 1
MA LX
1 0 0 0 0 0
1 1 0 0 0 0
1 1 1 0 0 0
1 1 1 1 0 0
1 1 1 1 1 0
1 1 1 1 1 1
OU ND = 6
```

Here we assume that  $\Sigma$  is in file **SIGMA**. We use ML to fit  $TT'$  to  $\Sigma$ , but any fit function in LISREL will do. The sample size on the DA command is arbitrary, but something larger than zero must be specified. Just make sure the “model” fits perfectly. All residuals and chi-square must be zero. The matrix  $T$  is given as LAMBDA-X in the output file:

LAMBDA-X

	KSI 1	KSI 2	KSI 3	KSI 4	KSI 5	KSI 6
VAR 1	1.000000	- -	- -	- -	- -	- -
VAR 2	0.378000	0.925806	- -	- -	- -	- -
VAR 3	0.720000	0.068956	0.690540	- -	- -	- -
VAR 4	0.324000	0.321372	0.047151	0.888550	- -	- -
VAR 5	0.270000	0.267810	0.039292	0.140229	0.913329	- -
VAR 6	0.270000	0.025858	0.063453	0.010374	0.006818	0.960339

### 3.2 Generating normal variables

The following PRELIS input file (**SIMEX11.PRL**) generates 200 independent cases of six variables having a multivariate normal distribution with zero mean vector and covariance matrix  $\Sigma$ . Note that there is no NI value specified on the DA command.

Generating Multivariate Normal Variables  
with a Specified Covariance Matrix

```
DA NO=200
NE V1=NRAND
NE V2=NRAND
NE V3=NRAND
NE V4=NRAND
NE V5=NRAND
NE V6=NRAND
```

```

NE X1=V1
NE X2=.378*V1+.925806*V2
NE X3=.72*V1+.068956*V2+.690540*V3
NE X4=.324*V1+.321372*V2+.047151*V3+.88855*V4
NE X5=.27*V1+.26781*V2+.039292*V3+.140229*V4+.913329*V5
NE X6=.27*V1+.025858*V2+.063453*V3+.010374*V4+.006818*V5+.960339*V6
CO ALL
SD V1-V6
OU RA=RAWDATA WI=7 ND=3 XM IX=123456

```

Here we generate data on V1 through V6 which are independent and normally distributed with mean zero and variance one. V1 through V6 are then transformed linearly to X1 through X6 using the **T** matrix. After X1 through X6 have been defined, V1 through V6 can be deleted. This is done with the SD command.

The raw data is obtained in the file **RAWDATA** in the format 6F7.3, WI = 7 defines the width of each field and ND = 3 defines the number of decimals for the data in the file **RAWDATA**.

Now, suppose we are not interested in the sample of raw data but only in the sample covariance matrix. Then, just replace the RA = RAWDATA specification with CM = SIMEX1.CM (see **SIMEX12.PRL**). Then no raw data will be saved or stored. The covariance matrix will be computed “on the run”.

Next suppose we want to generate 400 samples of size 200 and save all the 400 sample covariance matrices. Then, just add RP = 400 on the DA command, see file **SIMEX13.PRL**. The file **SIMEX1.CM** will then contain 400 sample covariance matrices. The first two replicates in **SIMEX1.CM** are:

```

0.95375D+00  0.37841D+00  0.10018D+01  0.59348D+00  0.30790D+00  0.83902D+00
0.36562D+00  0.40900D+00  0.32791D+00  0.11061D+01  0.20653D+00  0.30234D+00
0.19845D+00  0.28648D+00  0.87716D+00  0.22175D+00 -0.78280D-02  0.14026D+00
-0.12499D-01  0.32627D-01  0.10423D+01
0.87794D+00  0.37148D+00  0.12746D+01  0.60429D+00  0.41321D+00  0.86589D+00
0.29965D+00  0.49782D+00  0.28989D+00  0.92703D+00  0.14368D+00  0.39918D+00
0.15315D+00  0.26849D+00  0.98916D+00  0.30765D+00  0.17819D+00  0.25911D+00
0.30325D-01  0.40118D-01  0.11179D+01

```

This file can be read by LISREL to estimate a model for each sample. The following input file (**SIMEX14.LIS**) will estimate a confirmatory factor analysis model with two correlated factors, see Jöreskog (1979), Hägglund (1982), or Jöreskog & Sörbom (1989b, p.212)

```

Fitting a confirmatory factor analysis model in each of 400 samples
DA NI=6 NO=200 RP=400
CM=SIMEX1.CM
MO NX=6 NK=2
FR LX(1,1) LX(2,2) LX(3,1) LX(4,2) LX(5,2) LX(6,1)
OU AD=OFF PV=SIMEX1.PV SV=SIMEX1.SV GF=SIMEX1.GF XM

```

There are 13 free parameters in the model. The file **SIMEX1.PV** contains 400 sets of 13 estimated parameters, one from each sample. The file **SIMEX1.SV** contains 400 sets of 13 estimated standard errors of these parameter estimates, one from each sample. The file **SIMEX1.GF** contains 400 sets of all the 42 fit measures that LISREL computes for each sample. (Note that if the option XI is present on the OU command, LISREL will restrict the fit measures to the degrees of freedom, chi-square, and P-value). The information in each of these files may be studied further. Here we show how the parameter estimates in the file **SIMEX1.PV** may be examined.

The first two sets of parameter estimates in **SIMEX1.PV** are:

```

  1  0  0
0.853458D+00  0.660581D+00  0.698258D+00  0.641822D+00  0.433952D+00  0.224460D+00
0.655037D+00  0.225359D+00  0.565433D+00  0.351457D+00  0.694165D+00  0.688845D+00
0.991918D+00
  2  0  0
0.781581D+00  0.863154D+00  0.774720D+00  0.592991D+00  0.438305D+00  0.357914D+00
0.578626D+00  0.267072D+00  0.529566D+00  0.265699D+00  0.575391D+00  0.797048D+00
0.989797D+00

```

The first number is the replication number, the second and third are error indicators. The second number is

- 0** if iterations have converged and the  $P$ -value for  $\chi^2$  is in the interval  $.0005 \leq P \leq .995$
- 1** if iterations have not converged
- 2** if iterations have converged and the  $P$ -value for  $\chi^2$  is either  $P < .0005$  or  $P > .9995$ .

The third number is 0 if the solution is admissible, otherwise it is 1.

The following PRELIS input file (**SIMEX15.PRL**) analyzes the parameter estimates in file **SIMEX1.PV**.

```

Analyzing the parameter estimates in SIMEX1.PV
DA NI=14
LA
IND 'LX(1,1)' 'LX(2,2)' 'LX(3,1)' 'LX(4,2)' 'LX(5,2)' 'LX(6,1)'
      'PH(1,1)' 'TD(1)' 'TD(2)' 'TD(3)' 'TD(4)' 'TD(5)' 'TD(6)'
RA=SIMEX1.PV FO;(3X,F3.0/(6D14.6))
SD 1
CO 'LX(1,1)' - 'TD(6)'
OU

```

The output gives the following information about the sampling distribution of the parameter estimates.

#### Univariate Summary Statistics for Continuous Variables

Variable	Mean	St. Dev.	Skewness	Kurtosis	Minimum	Freq.	Maximum	Freq.
LX(1,1)	0.903	0.076	0.044	-0.305	0.705	1	1.136	1
LX(2,2)	0.697	0.085	0.115	0.091	0.485	1	0.956	1
LX(3,1)	0.801	0.074	-0.159	-0.046	0.560	1	0.990	1
LX(4,2)	0.602	0.079	-0.186	0.256	0.312	1	0.839	1
LX(5,2)	0.499	0.080	-0.141	-0.024	0.233	1	0.715	1
LX(6,1)	0.299	0.072	-0.066	0.059	0.089	1	0.489	1
PH(1,1)	0.606	0.077	-0.346	0.248	0.327	1	0.796	1
TD(1)	0.181	0.091	-0.315	-0.092	-0.075	1	0.401	1
TD(2)	0.511	0.097	-0.175	0.560	0.151	1	0.814	1
TD(3)	0.356	0.078	-0.208	0.726	0.053	1	0.577	1
TD(4)	0.627	0.087	0.115	0.070	0.382	1	0.902	1
TD(5)	0.742	0.087	0.115	-0.098	0.437	1	0.968	1
TD(6)	0.901	0.095	0.419	0.473	0.667	1	1.308	1

## Test of Univariate Normality for Continuous Variables

Variable	Skewness		Kurtosis		Skewness and Kurtosis	
	Z-Score	P-Value	Z-Score	P-Value	Chi-Square	P-Value
LX(1,1)	0.366	0.714	-1.385	0.166	2.054	0.358
LX(2,2)	0.949	0.343	0.479	0.632	1.129	0.569
LX(3,1)	-1.308	0.191	-0.086	0.932	1.717	0.424
LX(4,2)	-1.526	0.127	1.073	0.283	3.478	0.176
LX(5,2)	-1.159	0.247	0.010	0.992	1.342	0.511
LX(6,1)	-0.546	0.585	0.355	0.723	0.424	0.809
PH(1,1)	-2.793	0.005	1.046	0.295	8.897	0.012
TD(1)	-2.550	0.011	-0.292	0.770	6.587	0.037
TD(2)	-1.437	0.151	1.983	0.047	5.998	0.050
TD(3)	-1.704	0.088	2.406	0.016	8.689	0.013
TD(4)	0.948	0.343	0.396	0.692	1.056	0.590
TD(5)	0.953	0.340	-0.315	0.752	1.008	0.604
TD(6)	3.338	0.001	1.741	0.082	14.175	0.001

This reveals that

- The mean of the sampling distribution is close to the true population value for most parameters.
- There are some negative parameter estimates of the error variance TD(1) (Heywood cases), see comment in Jöreskog & Sörbom (1989b, p.212).
- The asymptotic normality approximation, which LISREL assumes when estimating standard errors of parameter estimates, may not be sufficiently close for some of the parameters.

This suggest that the sample size should be larger than 200 to avoid Heywood cases and to obtain correct standard errors.

### 3.3 Generating non-normal variables: Monte Carlo example 2

There are many non-normal variables and many ways of generating them.

The following example generates six variables  $x_1, \dots, x_6$  having the same population covariance matrix as in the previous example. These six variables are generated from  $v_1, \dots, v_6$ , which are independent, where

$v_1$  is normal with mean 0 and variance 1

$v_2$  is chi-square with 3 degrees of freedom

$v_3$  is a three-point distribution with probabilities  $\frac{1}{2}, \frac{1}{3}, \frac{1}{6}$  at 0, 1, and 2, respectively

$v_4$  is a uniform discrete distribution with probability  $\frac{1}{6}$  at 1, 2, 3, ..., 6

$v_5$  is uniform over the interval (0,1)

$v_6$  is  $w^2$  where  $w$  is uniform over the interval (0,1)

These variables are generated first, then they are standardized to zero mean and unit variance using the formula

$$v^* = \frac{1}{\sigma} v - \frac{\mu}{\sigma},$$

where  $\mu$  and  $\sigma$  are the mean and standard deviation of  $v$ , and  $v^*$  replaces  $v$ . Finally, they are transformed by the same matrix  $\mathbf{T}$  as before. The input file **SIMEX21.PRL** is:

```

Generating Non-Normal Variables
with a Specified Covariance Matrix
DA NO=200
NE V1=NRAND
NE V2=NRAND**2+NRAND**2+NRAND**2
NE V3=URAND
RE V3 OLD=0-.5,.50001-.83333,.83334-1 NEW=0,1,2
NE V4=URAND
RE V4 OLD=0-.16666,.16667-.33333,.33334-.5 NEW=1,2,3
RE V4 OLD=.50001-.66666,.66667-.83333,.83334-.99999 NEW=4,5,6
NE V5=URAND
NE V6=URAND**2
NE V2=.408248*V2-1.224745
NE V3=1.34164*V3-.894428
NE V4=.58554*V4-2.04939
NE V5=3.4641*V5-1.73205
NE V6=3.3541*V6-1.118033
NE X1=V1
NE X2=.378*V1+.925806*V2
NE X3=.72*V1+.068956*V2+.690540*V3
NE X4=.324*V1+.321372*V2+.047151*V3+.88855*V4
NE X5=.27*V1+.26781*V2+.039292*V3+.140229*V4+.913329*V5
NE X6=.27*V1+.025858*V2+.063453*V3+.010374*V4+.006818*V5+.960339*V6
CO ALL
SD V1-V6
OU CM=SIMEX2.CM XM IX=123

```

This way of generating  $x_1, \dots, x_6$  is not ideal because they are linear combinations of  $v_1, \dots, v_6$  and it is therefore, in general, difficult to know what characteristics they have apart from first and second order moments.

Using NRAND and linear combinations of its powers up to third order and methods developed by Fleishman (1978) and Vale & Maurelli (1983), it is possible to generate variables with specified univariate skewness and kurtosis and a specified covariance matrix. This method too has a disadvantage, for -unless the sample size is huge – there will be very large random variations in sample skewnesses and kurtoses from sample to sample. A better way may be to generate the latent and error variables and then generate the observable variables according to the LISREL model. This will be considered in the next section.

### 3.4 Generating variables from a specified model

Suppose a LISREL model to be simulated is specified. Suppose the model is recursive. Data on the observed variables  $y_1, \dots, y_p, x_1, \dots, x_q$  can be generated as follows

Step 1      Generate values on  $\xi_1, \dots, \xi_n$  and  $\zeta_1, \dots, \zeta_m$ .

- Step 2      Generate values on  $\varepsilon_1, \dots, \varepsilon_p$ .
- Step 3      Generate values on  $\delta_1, \dots, \delta_p$ .
- Step 4      Generate values on  $\eta_1, \dots, \eta_m$  from the structural equations in the LISREL model.
- Step 5      Generate values on  $y_1, \dots, y_p$  from the measurement model for the  $y$ -variables in the LISREL model.
- Step 6      Generate values on  $x_1, \dots, x_p$  from the measurement model for the  $x$ -variables in the LISREL model.

For notation and formulas, see Jöreskog & Sörbom (1989, p.4). If the model is non-recursive, the reduced form equations must be used instead of the structural equations in Step 4. For a LISREL submodel 1, as illustrated below, only the first part of Step 1 and Steps 3 and 6 are necessary.

### 3.5 Monte Carlo example 3

Suppose we want to simulate the following confirmatory factor analysis model with three factors:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} \lambda_1 & \lambda_2 & \lambda_3 \\ 1 & 0 & 0 \\ \lambda_4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_5 & 0 \\ \lambda_6 & \lambda_7 & 1 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \end{bmatrix},$$

where the covariance matrix of  $[\xi_1, \xi_2, \xi_3]$  is

$$\Phi = \begin{bmatrix} \phi_{11} & & \\ \phi_{21} & \phi_{22} & \\ 0 & 0 & \phi_{33} \end{bmatrix},$$

and the covariance matrix of  $[\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6]$  is

$$\Theta_\delta = \text{diag}[\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6].$$

We assume that

- $\xi_1, \xi_2$  and  $\xi_3$  are trivariate normal with zero means
- $\delta_i$  is uniformly distributed over the interval  $(0, \theta_i)$  where  $i = 1, 2, \dots, 6$
- $\delta_i$  is independent of  $\delta_j$  for  $i \neq j$
- $\delta_i$  is independent of  $\xi_j$  for  $i=1, 2, \dots, 6$  and  $j = 1, 2, 3$

and that the parameter values are

$$\begin{aligned}(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7) &= (.2, .4, .7, .7, .8, .2, .3) \\(\phi_{11}, \phi_{12}, \phi_{22}, \phi_{33}) &= (.64, .32, .65, .81) \\(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) &= (2, 2.5, 3, 3.5, 4, 4.5).\end{aligned}$$

Then data on the observed  $x$ -variables can be generated by the following PRELIS input file (**SIMEX31.PRL**):

```
Generating Sample Covariance Matrix for Confirmatory Factor Analysis Model
DA NO=200
CO ALL
NE KSI1=.7*NRAND
NE KSI2=.5*KSI1+.7*NRAND
NE KSI3=.9*NRAND
NE DELTA1=2*URAND
NE DELTA2=2.5*URAND
NE DELTA3=3*URAND
NE DELTA4=3.5*URAND
NE DELTA5=4*URAND
NE DELTA6=4.5*URAND
NE X1=.2*KSI1+.4*KSI2+.7*KSI3+DELTA1
NE X2=KSI1+DELTA2
NE X3=.7*KSI1+DELTA3
NE X4=KSI2+DELTA4
NE X5=.8*KSI2+DELTA5
NE X6=.2*KSI1+.3*KSI2+KSI3+DELTA6
SD KSI1-KSI3 DELTA1-DELTA6
OU CM=SIMEX3.CM XM IX=123
```

To obtain  $r$  replicates, just add  $RP = r$  on the DA command.

Several other variations of this are possible. For example,

- one can let both  $\xi$  and  $\delta$  be non-normal
- one can let some  $\xi$ -variable correlate with some  $\delta$ -variable
- one can let a  $\xi$ -variable and a  $\delta$ -variable be uncorrelated but functionally related

In the last two cases, fundamental assumptions of the LISREL model are violated, but it may be interesting to study what happens in such cases.

### 3.6 Generating ordinal variables: Monte Carlo example 4

We want to study the distribution of the goodness-of-fit statistics when a specified model (the same as in the bootstrap example) is estimated by WLS on the basis of polychoric correlations for ordinal variables. To make valid conclusions we want to simulate ordinal variables which are like the empirical data in **efficacy.raw** as much as possible.

#### Step 1

We begin by running PRELIS on **efficacy.raw** to estimate the marginal frequencies of each ordinal variable and the polychoric correlation matrix and its asymptotic covariance matrix.

The input file (**SIMEX41.PRL**) is

```
Computing Thresholds and PM and AC Matrices from EFFICACY.RAW
DA NI=6
LA=EFFICACY.LAB
RA=EFFICACY.RAW FO;(6F1.0)
OU MA=PM PM=EFFICACY.PML AC=EFFICACY.ACP TH=EFFICACY.TH
```

The thresholds are reported in the output file **efficacy.th**:

```
-1.522  0.243  1.221
-1.876 -0.367  0.908
-1.746 -0.799  0.656
-1.829 -0.165  1.074
-2.124 -0.545  0.974
-1.985 -0.349  1.030
```

## Step 2

Next, we fit the model to the matrix of polychoric correlations with the WLS method. Thus, run LISREL with the following command file (**SIMEX42.LIS**):

```
Computing and Saving Sigma for Measurement Model
DA NI=6 NO=297 MA=PM
LA=EFFICACY.LAB
PM=EFFICACY.PML
AC=EFFICACY.ACP
MO NX=6 NK=2 PH=ST
FR LX(1,1) LX(2,1) LX(3,1) LX(4,1) LX(4,2) LX(5,2) LX(6,2)
OU SI=EFFICACY.SIG ME=WLS
```

The fitted covariance matrix in file **efficacy.sig** will be used together with the thresholds obtained in Step 1 to generate the random data in the next step. Since the factor loadings obtained in this run fit the correlation matrix in file **efficacy.sig** perfectly, this means that the model we will simulate holds exactly but only approximately. For this purpose, use the sample correlation matrix **efficacy.pml** obtained in Step 1 instead of **efficacy.sig** in what follows.

## Step 3

To generate ordinal variables with the marginal distributions obtained in Step 1, we need a lower triangular matrix **T** that factorizes the matrix in **efficacy.sig**. This is obtained as in Monte Carlo example 1. The matrix **T** is:

```
1.000000
.313390 .950978
.376890 .173843 .909798
.490220 .226112 .242650 .806026
.379350 .174984 .187775 .475599 .751013
.417710 .192668 .206755 .523676 .204797 .655341
```

We can then generate the ordinal variables and their polychoric correlations for each sample without actually storing the raw data. We can also obtain the estimated asymptotic covariance matrix of the polychoric correlations for each sample. All this is done by running the following input file (**SIMEX43.PRL**):

```

Estimating 200 polychoric correlation matrices from Monte Carlo data
DA NO=400 RP=200
NE V1=NRAND
NE V2=NRAND
NE V3=NRAND
NE V4=NRAND
NE V5=NRAND
NE V6=NRAND

NE X1=V1
NE X2=.311390*V1+.950282*V2
NE X3=.376890*V1+.173843*V2+.909798*V3
NE X4=.490220*V1+.226112*V2+.242650*V3+.806026*V4
NE X5=.379350*V1+.174984*V2+.187775*V3+.475599*V4+.751013*V5
NE X6=.417710*V1+.192668*V2+.206755*V3+.523676*V4+.204797*V5+.655341*V6

RE X1 OLD=-20--1.522,-1.522-.243,.243-1.221,1.221-20 NEW=1,2,3,4
RE X2 OLD=-20--1.876,-1.876--.367,-.367-.908,.908-20 NEW=1,2,3,4
RE X3 OLD=-20--1.746,-1.746--.799,-.799-.656,.656-20 NEW=1,2,3,4
RE X4 OLD=-20--1.829,-1.829--.165,-.165-1.074,1.074-20 NEW=1,2,3,4
RE X5 OLD=-20--2.124,-2.124--.545,-.545-.974,.974-20 NEW=1,2,3,4
RE X6 OLD=-20--1.985,-1.985--.349,-.349-1.030,1.030-20 NEW=1,2,3,4
SD V1-V6
OU IX=2345 MA=PM PM=EFFICACY.PMM AC=EFFICACY.ACM XB XT

```

The first part of this input generates independent random normal deviates. The second part transforms these, using the **T** matrix, to a set of multinormal variables with covariance matrix **efficacy.sig**. Finally, in the last part, these multinormal variables are ordinalized by grouping them in intervals according to the thresholds in **efficacy.th**.

After this run, we have 200 matrices of polychoric correlations in the file **efficacy.pmm** and 200 corresponding asymptotic covariance matrices in the file **efficacy.acm**.

Alternatively, we could use the fitted solution from Step 2 to generate  $\xi_1, \xi_2, \delta_1, \dots, \delta_6$  first and then the underlying  $x$ -variables as linear combinations of these according to the model, as in Monte Carlo example 3. The ordinalization remains the same. We leave it to the reader to write the input file for the alternative.

#### Step 4

Each of these matrices will now be analyzed by fitting a LISREL model to the polychoric correlations by WLS. For each sample analyzed, all the goodness-of-fit measures will be saved in the file **efficacy.gfm**. The input file (**simex44.lis**) is:

```

Generating 200 Sets of Fit Measures from EFFICACY.PMM and EFFICACY.ACM
DA NI=6 NO=400 MA=PM RP=200
LA=EFFICACY.LAB REWIND
PM=EFFICACY.PMM
AC=EFFICACY.ACM
MO NX=6 NK=2 PH=ST
FR LX(1,1) LX(2,1) LX(3,1) LX(4,1) LX(4,2) LX(5,2) LX(6,2)
OU GF=EFFICACY.GFM ME=WLS

```

The fit measures for the first Monte Carlo sample look like this:

```

1 0 0 7 0.58793D+01 0.55391D+00 0.00000D+00 0.77601D-01 0.00000D+00
0.10000D+01 0.00000D+00 0.10000D+01 0.00000D+00 0.00000D+00 0.85375D+01
0.14735D-01 0.00000D+00 0.00000D+00 0.21397D-01 0.00000D+00 0.00000D+00
0.55288D-01 0.92288D+00 0.87719D-01 0.87719D-01 0.10912D+00 0.10526D+00
0.27592D+01 0.00000D+00 0.00000D+00 0.00000D+00 0.00000D+00 0.00000D+00
0.00000D+00 0.00000D+00 0.18658D-01 0.18658D-01 0.99831D+00 0.99494D+00
0.33277D+00 0.99460D+00 0.10022D+01 0.46415D+00 0.10000D+01 0.10010D+01
0.98843D+00 0.00000D+00

```

The first number is the replication number. The next two numbers are the same error indicators described earlier. The fourth integer is the degrees of freedom for the model. The other numbers are all 43 fit measures that LISREL computes and they are given in exactly the same order they appear in the output file. For a definition of the fit measures, see Jöreskog & Sörbom (1993). We note that chi-square is the first measures, the *P*-value is the second, and RMSEA is the 15<sup>th</sup>. The numbers following the degrees of freedom are the four chi-squares and their corresponding *P*-values.

## Step 5

We can now analyze the 200 sets of fit measures in the **efficacy.gfm** file. Here we study only the three measures mentioned previously. The input file (**SIMEX45.PRL**) is:

```

Analyzing 200 Sets of Fit Measures in EFFICACY.GFM
DA NI=3
LA;
CHI2 P RMSEA
RA=EFFICACY.GFM FO;(16X,2D13.5//52X,D13.5/////);
CO ALL
OU MA=KM WP

```

The following results are obtained.

Total Sample Size(N) = 200

### Univariate Summary Statistics for Continuous Variables

Variable	Mean	St. Dev.	Skewness	Kurtosis	Minimum	Freq.	Maximum	Freq.
CHI2	7.610	4.304	1.614	4.877	0.696	1	30.528	1
P	0.461	0.289	0.143	-1.159	0.000	1	0.998	1
RMSEA	0.016	0.020	1.075	0.423	0.000	107	0.092	1

### Test of Univariate Normality for Continuous Variables

Variable	Skewness		Kurtosis		Skewness and Kurtosis	
	Z-Score	P-Value	Z-Score	P-Value	Chi-Square	P-Value
CHI2	7.097	0.000	5.339	0.000	78.874	0.000
P	0.846	0.397	-8.619	0.000	75.004	0.000
RMSEA	5.353	0.000	1.228	0.219	30.163	0.000

Relative Multivariate Kurtosis = 3.693

Test of Multivariate Normality for Continuous Variables

Skewness			Kurtosis			Skewness and Kurtosis	
Value	Z-Score	P-Value	Value	Z-Score	P-Value	Chi-Square	P-Value
38.451	27.263	0.000	55.388	11.129	0.000	867.143	0.000

Correlation Matrix

	CHI2	P	RMSEA
CHI2	1.000		
P	-0.894	1.000	
RMSEA	0.942	-0.869	1.000

Total Variance = 3.000 Generalized Variance = 0.0218

Largest Eigenvalue = 2.804 Smallest Eigenvalue = 0.055

Condition Number = 7.137

Means

CHI2	P	RMSEA
7.610	0.461	0.016

Standard Deviations

CHI2	P	RMSEA
4.304	0.289	0.020

With 7 degrees of freedom the theoretical mean of chi-square is 7 and the standard deviation is  $\sqrt{14} = 3.742$ . It is seen that the standard deviation is underestimated. All three fit measures are almost perfectly correlated. There are 176 zero values of RMSEA. This is the number of times out of 200 that chi-square is below the degrees of freedom.

We can test the hypothesis that the chi-square is roughly distributed as  $\chi^2$  with 7 degrees of freedom. This can be tested by grouping the chi-square measure into 10 equally probable intervals, as done with the following input file (**SIMEX46.PRL**):

```
Analyzing 200 Sets of Chi-square Measures in EFFICACY.GFM
DA NI=1
LA
CHI2
RA=EFFICACY.GFM FO;(16X,D13.5////////)
RE CHI2 OLD=0-2.83,2.8301-3.82,3.8201-4.67,4.6701-5.49 NEW=1,2,3,4
RE CHI2 OLD=5.4901-6.35,6.3501-7.28,7.2801-8.38,8.3801-9.80 NEW=5,6,7,8
RE CHI2 OLD=9.801-12.0,12.001-4711 NEW=9,10
OU
```

The output file gives the following distribution for the grouped chi-square.

CHI2	Frequency	Percentage	Bar Chart
1	17	8.5	██
2	16	8.0	██
3	15	7.5	██
4	21	10.5	██
5	23	11.5	██
6	19	9.5	██
7	18	9.0	██
8	25	12.5	██
9	20	10.0	██
10	26	13.0	██

It is quite obvious from this that chi-square follows a  $\chi^2$  with 7 degrees of freedom rather closely. A formal chi-square test of this gives  $\chi^2 = 6.7$  with 9 degrees of freedom.

### 4. References

Fleishman, A.I. (1978) A method for simulating non-normal distributions. *Psychometrika*, **43**, 521-532.

Hägglund, G. (1982) Factor analysis by instrumental variable methods. *Psychometrika*, **47**, 209-222.

Jöreskog, K.G. (1979) Basic ideas of factor and component analysis. In: K.G. Jöreskog & D. Sörbom: *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books, 5-20.

Jöreskog, K.G. & Sörbom, D. (1989) *LISREL 7: User's Reference Guide*. Chicago, IL: Scientific Software International.

Jöreskog, K.G. & Sörbom, D. (1989b; 2nd edition). *LISREL 7: A guide to the Program and Applications*. Chicago, IL: Scientific Software International.

Jöreskog, K.G. & Sörbom, D. (1993) *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*. Chicago, IL: Scientific Software International.

Schrage, L. (1979) A more portable FORTRAN random number generator. *ACM Transactions on Mathematical Software*, **5**, 132-138.

Vale, C.D. & Maurelli, V.A. (1983) Simulating multivariate nonnormal distributions. *Psychometrika*, **48**, 465-471.