Before LISREL: Preparing the Data using PRELIS

The data used in a linear structural equation model are the sufficient statistics (means, variances and covariances).

Before any analysis this statistics have to be computed. This can be done with the program PRELIS which is included in the Interactive LISREL system.

Interactive LISREL can read in an ordinary data matrix with rows for cases and columns for variables from many statistical programs. But in the student version this is restricted to read in only a SPSS system file (with no more than 20 variables) or ASCII data, where the realizations are separated by blanks (free format, expected extension is ".dat"), commas (expected extension is ".csv") or tabulator stops (expected extensions is ".txt").

The data of the efficacy example are stored in a SPSS system-file called "ALLB96SUB.SAV". Additionally to the seven variables polint1, polint2, impact, election, politicn, governm and leader the variable group is read in as first variables where the code "1" denotes respondents from the old (Western) states and the code "2" denotes respondents from the new (Eastern) states of Germany. The file can be read in from the menu.
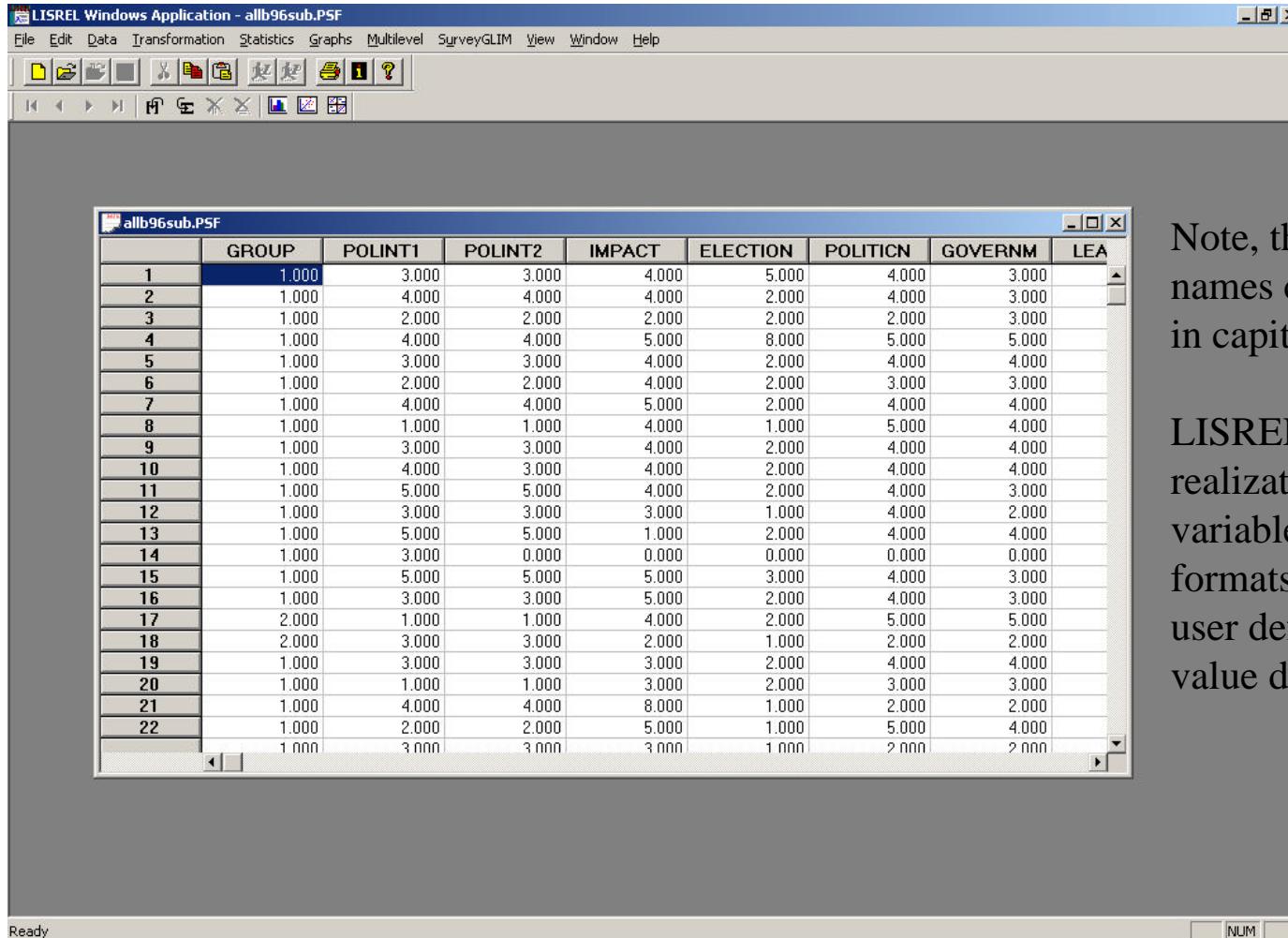
Reading SPSS system files in LISREL

To read in: File → Import Data in Free Format → file form: SPSS for Windows
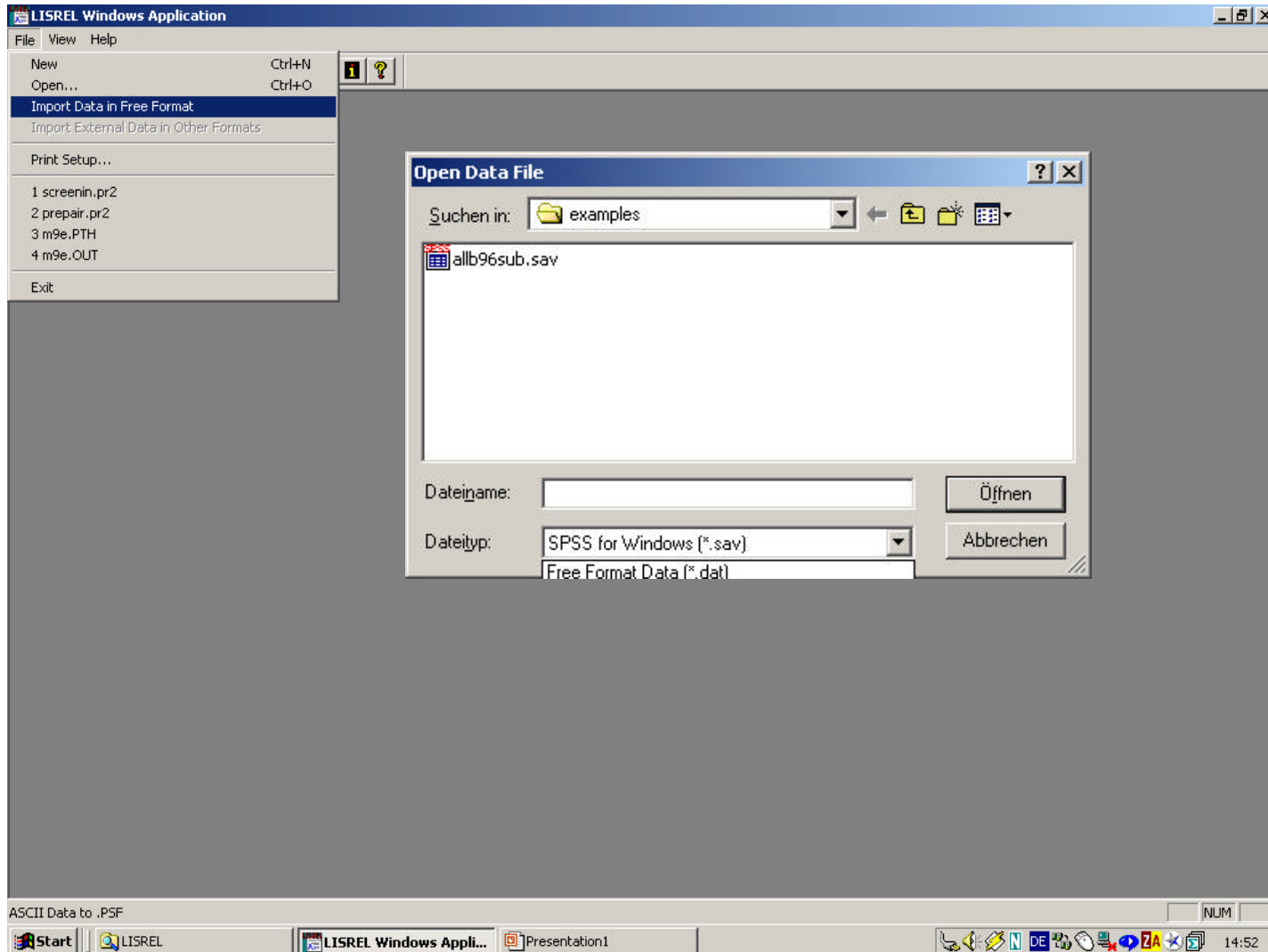
Reading SPSS system-files in LISREL

The result will be a spreadsheet like data set, that is called PRELIS system file and which name is the same as the SPSS system-file but the extension ".PSF", here "allb96sub.PSF".

Note, that the variable names come from SPSS in capital letters.

LISREL reads only the realizations, but ignores variable or value labels, formats, scale levels and user defined missing value declarations.

Reading ASCII files with fixed formats

Using PRELIS-commands it is also possible to read ASCII files where the variables are stored in fixed format.

For example, the data allb96sub.dat includes the same data as the SPSS system-file.

The first and last two rows are:

```
13345431
14442434

  . . .

11132226
14433339
```

There is no space between the numbers. The FORTRAN input format for these data is (8F1.0) where "F1.0" means that a variable is stored in one column and that there are no implicit decimals. This format is reiterated without any space 8 times for the 8 variables in the data set.

In general:

kFn.m means: k variables stored in n columns each where the last m columns are digits after the decimals point if there is no decimal point given.

kAn   means: k stings (for example labels) are stored where each string has a length of n characters

kX   means: skip the next k columns

/   means: skip the rest of the line and go to next line.

Use commas to separate elements, example (3X,2F2.0,1X,/,8F1.0)   4

Reading ASCII files with fixed formats

PRELIS commands are given in any ascii-file (preferred extension is ".PR2"). The command for reading the data file "allb96sub.dat" are:

line	PRELIS command
01	Read in formatted raw data
02	DA NI=8
03	LA
04	Group Polint1 Polint2 Impact
05	Election Politicn Governm
06	Leader /
07	RA=allb96sub.dat FO
08	(8F1.0)
09	SD Group=1
10	OU RA=allb96sub.psf

line	description
01	optional title
02	DA-command gives number of input variable NI (in the example ni=8)
03	optional LA-command forces PRELIS to read in NI labels (variable names) for the variables read in LA=filename if labels are stored in a file

line	description
04-06	variable labels (note: no more than 8 characters each, the names are case sensitive) the ending slash ("/") denotes the end of the labels (if one does not give labels for all variables
07	RA-command forces PRELIS to read in raw data from a file or from the following lines. PRELIS expected either free formated data (blanks between each realization) or a FORTRAN format in the first line(s) of the data file. The options FO means that the FORTRAN format comes on the next line
08	FORTRAN format for the input data must be in bracket

Reading ASCII files with fixed formats

line	PRELIS command
01	Read in formatted raw data
02	DA NI=8
03	LA
04	Group Polint1 Polint2 Impact
05	Election Politicn Governm
06	Leader /
07	RA=allb96sub.dat FO
08	(8F1.0)
09	SD Group=1
10	OU RA=allb96sub.psf

line	description
09	SD-command select cases and delete the selection variable. After the command follows the variables used for selection and then a condition: = value, > value, < value
10	output command the option RA forces to save the data as a raw data in the file given. If as in the example the extension of the raw data set is .psf a psf-file will be generated.

Further options of the ou-command for writing raw data files

If raw data should be written out as ascii files there are two options to format them:

WI=n   n columns will be used for each variable

ND=m   m digits after the decimal are printed.

for example the data in the raw data file read in can be written out again by the PRELIS command: ou ra=allb96sub.dat wi=1 nd=0

Preparing the raw data

The variables in the SPSS file have a different order and different missing values:

The missing value codes are:
0: not asked, 8: don't know, 9: no answer

Polint1, Polint2, Impact, Election, Politicn and Governm are coded so, that the lowest value (1) means high political interest or agree strongly to an item
whereas the highest value (5) means no political interest or disagree strongly to an item.

For the interpretation of the results of a structural equation model it is easier if the coding of the variables follows its meaning. Otherwise there is the possibility of false interpretation of effect signs.

Therefore, after reading in and looking at the variable distribution, the data should be repoled. Additionally all missing value codes will be recoded in missing value code 9.

Preparing the raw data

PRELIS commands for recoding::

```
Preparing data
SY=allb96sub.psf
RE Polint1 OLD=1,2,3,4,5 NEW=5,4,3,2,1
RE Polint2 OLD=1,2,3,4,5,0,8-9 NEW=5,4,3,2,1,9,9
RE Impact OLD=1,2,3,4,5,0,8 NEW=5,4,3,2,1,9,9
RE Election OLD=1,2,3,4,5,0,8 NEW=5,4,3,2,1,9,9
RE Politicn OLD=1,2,3,4,5,0,8 NEW=5,4,3,2,1,9,9
RE Governm OLD=1,2,3,4,5,0,8 NEW=5,4,3,2,1,9,9
OU RA=allb96sub.psf
```

SY=filename read a Prelis system file

RE recode

old=... value (ranges) before recode

new=... recoded values

Example:

Original Data			Recoded Data		
Polint2	Frequency	Percentage	Polint2	Frequency	Percentage
0	41	1.7			
1	169	7.0	1	171	7.1
2	448	18.7	2	497	20.7
3	1024	42.6	3	1024	42.6
4	497	20.7	4	448	18.7
5	171	7.1	5	169	7.0
8	16	0.7			
9	36	1.5	9	93	3.9

Screening missing values pattern

If there missing values their pattern should be investigated:
```
Screening missing value pattern
SY=allb96sub.psf
MI 9 ALL
OU
```

MI   Missing values command
first missing values, than variables

```
Number of Missing Values per Variable
   Polint1    Polint2    Impact   Election  Politicn   Governm    Leader
   --------   --------   --------  --------  --------   --------   --------
        3         93        136       152       226        241        217


 Distribution of Missing Values
 Total Sample Size =    2402
 Number of Missing Values     0     1     2     3     4     5     6     7
        Number of Cases  1882   285    90    51    32    51    10     1
 Effective Sample Sizes
 Univariate (in Diagonal) and Pairwise Bivariate (off Diagonal)
              Polint1    Polint2     Impact    Election   Politicn    Governm
              --------   --------   --------   --------   --------   --------
  Polint1       2399
  Polint2       2307      2309
   Impact       2264      2228       2266
   ...
```

Screening missing values pattern

```
Percentage of Missing Values
 Univariate (in Diagonal) and Pairwise Bivariate (off Diagonal)
          Polint1     Polint2      Impact    Election    Politicn     Governm
         --------    --------    --------    --------    --------    --------
  Polint1     0.12
  Polint2     3.96        3.87
   Impact     5.75        7.24        5.66
 Election     6.41        7.74        8.28        6.33
 Politicn     9.49       10.91       10.70       10.87        9.41
  Governm    10.12       11.41       11.78       11.78       12.53       10.03
   Leader     9.12       12.32       13.32       13.99       15.86       16.74

            Leader
           --------
   Leader     9.03

Missing Data Map
Frequency PerCent    Pattern
     1882    78.4    0 0 0 0 0 0 0
        2     0.1    1 0 0 0 0 0 0
       27     1.1    0 1 0 0 0 0 0
       22     0.9    0 0 1 0 0 0 0

      ...
        7     0.3    0 0 1 1 1 1 1
       10     0.4    0 1 1 1 1 1 1
        1     0.0    1 1 1 1 1 1 1
```

10

Computation of Means, Covariances and Asymptotic Covariances

In the last step the sufficient statistics will be computed.
There are two aspects that have to be considered here:
- coping with missing values
- variable type

By default LISREL use listwise deletion for missing data.
But it is also possible to compute means, variances and covariances using pairwise deletion, impute missing values, or estimate means, variances and covariances by the EM algorithm.

By default, the variable type is ordinal (OR) as long as a variable has no more than 15 categories;
variables with 16 or more categories are metric continuously (CO) by default.
Other variable types are:
CA censored above, that is continuous but all values $\leq$ a number are recode to that number
CB censored below, that is continuous but all values $\geq$ a number are recode to that number
CE censored above and below

The PRELIS commands OR, CO, CA, CB or CE can be used to change the variable type.

Computation of Means, Covariances and Asymptotic Covariances

```
Computation of Means Covariances and Asymptotic Covariances
for West German Efficacy data (ALLBUS 1996) using listwise deletion
SY=allb96sub.psf
MI 9 all
CO all
OU MA=CM CM=a96wlstw.cm ME=a96wlstw.me AC=a96wlstw.acc PA
```

MI Missing value code 9 is defined for all variables

CO all variables are defined as metric

MA=type matrix to becompute will be

    CM for covariance matrix

    KM for correlation matrix (PM if ordinal data)

    AM for augmented moment matrix

CM=file stores the covariacne matrix in an ascii file

ME=file stores the means in an ascii file

AC=file estimate the covariance matrix of the sampling distribution of the covariances
       using 4th-order moments and store them in a file

PA   print all output.

Results

```
Total Sample Size =    2402
Number of Missing Values      0     1     2     3     4     5     6     7
         Number of Cases   1882   285    90    51    32    51    10     1
 Listwise Deletion
 Total Effective Sample Size =    1882
```

Univariate Summary Statistics for Continuous Variables

Variable	Mean	St. Dev.	T-Value	Skewness	Kurtosis	Minimum	Freq.	Maximum	Freq.
Polint1	2.894	0.975	128.732	-0.089	-0.180	1.000	176	5.000	88
Polint2	2.898	0.946	132.854	-0.074	-0.074	1.000	156	5.000	82
Impact	3.524	1.025	149.175	-0.383	-0.642	1.000	39	5.000	308
Election	2.279	1.039	95.156	0.855	0.163	1.000	393	5.000	70
Politicn	3.491	1.045	144.936	-0.291	-0.752	1.000	41	5.000	325
Governm	3.291	1.008	141.600	-0.003	-0.894	1.000	28	5.000	222
Leader	2.740	1.825	65.116	0.781	-0.476	1.000	715	7.000	93

Covariance Matrix

	Polint1	Polint2	Impact	Election	Politicn	Governm
Polint1	0.951					
Polint2	0.827	0.895				
Impact	0.098	0.118	1.050			
Election	0.128	0.135	0.296	1.079		
Politicn	0.044	0.054	0.284	0.238	1.092	
Governm	0.037	0.052	0.240	0.269	0.533	1.017
Leader	0.059	0.064	-0.194	-0.281	-0.292	-0.399

	Leader
Leader	3.331

Computing Covariances after pairwise deletion of missing values

```
Computing CM by pairwise deletion
DA NI=8 MI=0,8,9 TR=PA
LA
Group Polint1 Polint2 Impact
Election Politicn Governm Leader /
RA=allb96sub.dat FO
(8F1.0)
sd Group = 1
RE Polint1 OLD=1,2,3,4,5 NEW=5,4,3,2,1
RE Polint2 OLD=1,2,3,4,5 NEW=5,4,3,2,1
RE Impact OLD=1,2,3,4,5 NEW=5,4,3,2,1
RE Election OLD=1,2,3,4,5 NEW=5,4,3,2,1
RE Politicn OLD=1,2,3,4,5 NEW=5,4,3,2,1
RE Governm OLD=1,2,3,4,5 NEW=5,4,3,2,1
CO all
OU MA=CM
```

MI   Missing values can be defined also as option on the DA-command

TR=PA for parwise deletion.

	Polint1	Polint2	Impact	Election	Politicn	Governm
Polint1	1.050					
Polint2	0.909	0.998				
Impact	0.124	0.145	1.077			
Election	0.139	0.142	0.281	1.051		
Politicn	0.033	0.047	0.292	0.228	1.105	
Governm	0.023	0.039	0.237	0.259	0.546	1.029
Leader	-0.106	-0.090	0.158	0.273	0.287	0.423

	Leader
Leader	3.377

Data definition within the PSF file menue

All data definitions and transformations can be done also within the PSF-file editor:

EM-algorithm for coping with missing data

From a psf-file with defined global missing values a variance-covariance matrix covariance matrix can be estimated where missing data are imputed by the EM-algorithm:

```
Coping with Missing data using EM algorithm
SY=allb96sub.PSF
EM CC = 0.00001 IT = 200 TC = 2
OU MA=CM XT XM
```

EM   EM-command forces EM-algorithm for estimation means and covariances in data
     sets with missing data

cc=x   convergence criterion: estimations stops if change is less than x

IT=n   maximal number of iterations: Iterations stops after n oterations

TC=k   coping with cases where all variables are missing
     k=0: mean substitution (default value)
     k=1: case are ignored
     k=2: case will be deleted from the data set.

The results is:

```
---------------------------------
 EM Algoritm for missing Data:
---------------------------------
         Number of different missing-value patterns=       40
         Convergence of EM-algorithm in      4 iterations
         -2 Ln(L) =      43059.93590
         Percentage missing values=    6.31
```

EM-algorithm for coping with missing data

```
Estimated Means
 Polint1        2.9792
 Polint2        2.9642
 Impact         2.4650
 Election       3.7275
 Politicn       2.5364
 Governm        2.7261
 Leader         2.7832
```

```
          Estimated Covariances
              Polint1    Polint2    Impact     Election   Politicn   Governm
 Polint1       1.0497
 Polint2       0.9169     1.0056
 Impact        0.1229     0.1450     1.0763
 Election      0.1401     0.1479     0.2794     1.0514
 Politicn      0.0304     0.0423     0.2895     0.2247     1.1049
 Governm       0.0271     0.0417     0.2354     0.2541     0.5418     1.0255
 Leader       -0.1145    -0.1087     0.1586     0.2711     0.2982     0.4223
```

```
              Leader
 Leader        3.3803
```

```
Note: There are        1 cases with missing values
on all of the selected variables.
Cases with all values missing will be deleted.
```

Exercise

1) Produce raw data files from "allb96sub.dat" where the realizations are separated a) by blank, b) by comma and c) by tab stops and read in this data sets in a PSF-file using "import data in free format".

2) PRELIS has further methods to cope withg missing data.Describe this methods shortly
   (hint: using the LISREL help you can download technical papers on this issue)
   and compare the variancee-covariance matrix produced after coping with different methods of missing data handling.