Latent Variable Scores and Observational Residuals

Karl G Jöreskog
Dag Sörbom
Fan Yang Wallentin

Uppsala University

February 24, 2006

There are several methods available for estimating latent variable scores[1], *e.g.*, Lawley & Maxwell (1971, Chapter 8) or Bartholomew & Knott (1999, pp. 65–68). The two most commonly used methods for estimating latent variable scores are the regression method of Thomson (1939) and the Bartlett method (Bartlett, 1937). In LISREL we use the procedure of Anderson & Rubin (1956) as described in Jöreskog (2000). This procedure has the advantage of producing latent variable scores that have the same covariance matrix as the latent variables themselves.

In LISREL 8.8 we have added the possibility of estimating individual scores for all the error terms, measurement errors as well as structural errors, in any single-group LISREL model. Following Bollen & Arminger (1991) we use the general term *observational residuals* for this. The observational residuals depend on the method used for estimating the latent variable scores. Bollen & Arminger (1991) gave formulas which are valid for any method of estimating latent variable scores as linear combinations of observed variables. The results reported here are based on latent variable scores estimated by the Anderson & Rubin method.

This paper describes how latent variable scores and observational residuals can be obtained with LISREL 8.8 and illustrates their use with two examples.

1 Example 1: Interaction Model

The PRELIS system file **KJUDD.PSF** in the `NSFEX` subfolder contains data on the five variables y, x_1, x_2, x_3, x_4 generated according to the Kenny Judd model:

$$\begin{pmatrix} y \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} \alpha \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{pmatrix} + \begin{pmatrix} \gamma_1 & \gamma_2 & \gamma_3 \\ 1 & 0 & 0 \\ \lambda_2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_4 & 0 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_1\xi_2 \end{pmatrix} + \begin{pmatrix} \zeta \\ \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \end{pmatrix} \tag{1}$$

See Jöreskog & Yang (1996) for explanation of the terms and assumptions of this model. Note that y is a nonlinear function of the latent variables ξ_1 and ξ_2. This model was first considered by Kenny & Judd (1984) and several methods have been developed for estimating it, *e.g.*, Marsh, Wen, & Hau (2004) and references therein. Here we describe how the model can be estimated by

[1] In classical exploratory factor analysis these are usually called factor scores

using latent variable scores. At the same time we illustrate how observational residuals can be obtained.

We begin by estimating the linear part of the model, *i.e.*, the model without the interaction effect. This model is shown in the path diagram in Figure 1.

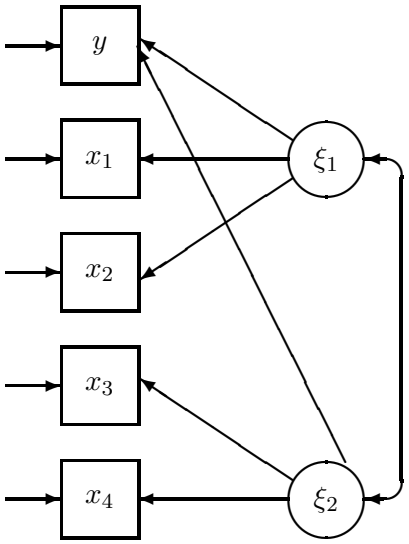

Figure 1: The Linear Part of the Kenny-Judd Model

The data is in the PRELIS system file **KJUDD.PSF**. See Jöreskog, *et al.* (2003, p. 168) for information about PSF files and how to create them. The following SIMPLIS command file **KJ1.SPL** estimates a standard linear confirmatory factor analysis model with two latent variables ξ_1 and ξ_2 (called `Ksi1` and `Ksi2` in the SIMPLIS file) and estimates the latent variable scores and all observational residuals at the same time.

```
Raw data from file KJUDD.PSF
Latent Variables: Ksi1 Ksi2
Relationships
Y=Ksi1 Ksi2
X1=1*Ksi1
X2=Ksi1
X3=1*Ksi2
X4=Ksi2
Path Diagram
PSFfile KJUDD.PSF
Estimate Residuals
End of Problem
```

In addition to the path diagram and the standard SIMPLIS output file **KJ1.OUT**, this run also generates a new PSF file **KJUDDnew.PSF** containing all the following variables

```
Y X1 X2 X3 X4 Ksi1 Ksi2 R_Y R_X1 R_X2 R_X3 R_X4
```

The variables

```
Ksi1 Ksi2
```

contain the estimated latent variable scores of ξ_1 and ξ_2. The variables

```
R_Y R_X1 R_X2 R_X3 R_X4
```

contain the estimated scores on $\zeta, \delta_1, \delta_2, \delta_3, \delta_4$. LISREL automatically puts `R_` in front of the variable name to mean the residual or error term. Thus, `R_varname` means the residual of or the error term on the variable `varname`. Since variable names in LISREL can contain at most 8 characters, it is advisable that `varname` contains at most 6 characters so that the extended variable `R_varname` will contain at most 8 characters.

In the file **KJ1.SPL** it is the line

```
PSFfile KJUDD.PSF
```

that produces the latent variable scores and the lines

```
PSFfile KJUDD.PSF
Estimate Residuals
```

that produce the observational residuals. Obviously, the observational residuals cannot be produced without the latent variable scores. Thus, one can omit the line

```
Estimate Residuals
```

if one is only interested in latent variable scores.

With the file **KJUDDnew.PSF** displayed, one can plot `Y` against `Ksi1` or `Ksi2`. These plots clearly shows that there are nonlinear relationships.

The product variable $\xi_1\xi_2$ can be computed and the nonlinear relationship between y and ξ_1 and ξ_2 can be estimated at the same time using the following PRELIS command file (**KJ2.PR2**)

```
Estimating the regression of Y on Ksi1, Ksi2 and Ksi1*Ksi2
SY=KJUDDnew.PSF
NE Ksi1Ksi2 = Ksi1*Ksi2
CA ALL
RG Y on Ksi1 Ksi2 Ksi1Ksi2
OU XU
```

The `XU` on the `OU` line is an option to skip the printing of univariate summary statistics.

The output **KJ2.OUT** gives the following estimated equation

```
       Y = 1.085  + 0.198*Ksi1 + 0.482*Ksi2 + 0.458*Ksi1Ksi2
 Standerr  (0.0174) (0.0246)     (0.0242)     (0.0263)
 Z-values   62.380   8.032        19.918       17.410
 P-values   0.000    0.000        0.000        0.000

            + Error,

 Error Variance = 0.271
```

which clearly shows that the estimate of γ_3 is highly significant.

Alternatively, one can use the estimated error term `R_Y` and regress this on `Ksi1Ksi2` only, using the following PRELIS syntax file (**KJ3.PR2**):

```
Estimating the regression of R_Y on Ksi1*Ksi2
SY=KJUDDnew.PSF
NE Ksi1Ksi2 = Ksi1*Ksi2
CO ALL
RG R_Y on Ksi1Ksi2
OU XU
```

This gives the following result

```
      R_Y =  - 0.0949 + 0.451*Ksi1Ksi2 + Error,
 Standerr     (0.0173) (0.0251)
 Z-values     -5.481    17.931
 P-values      0.000    0.000

 Error Variance = 0.271
```

Note that the two alternatives give the same estimates of the error variance and almost the same estimates of γ_3.

2 Example 2: Political Democracy

Bollen (1989, p. 17) presents a panel model of political democracy and industrialization in 75 countries. Bollen & Arminger (1991) used the same model in their discussion of observational residuals. The model is shown in the path diagram in Figure 2.

The variables in the model are

- y_1 Freedom of press 1960
- y_2 Freedom of political opposition 1960
- y_3 Fairness of elections 1960
- y_4 Effectiveness of legislature 1960
- y_5 Freedom of press 1965
- y_6 Freedom of political opposition 1965
- y_7 Fairness of elections 1965
- y_8 Effectiveness of legislature 1965
- x_1 GNP per capita 1960
- x_2 Energy consumtion per capita 1960
- x_3 Percentage of labor force in industry 1960

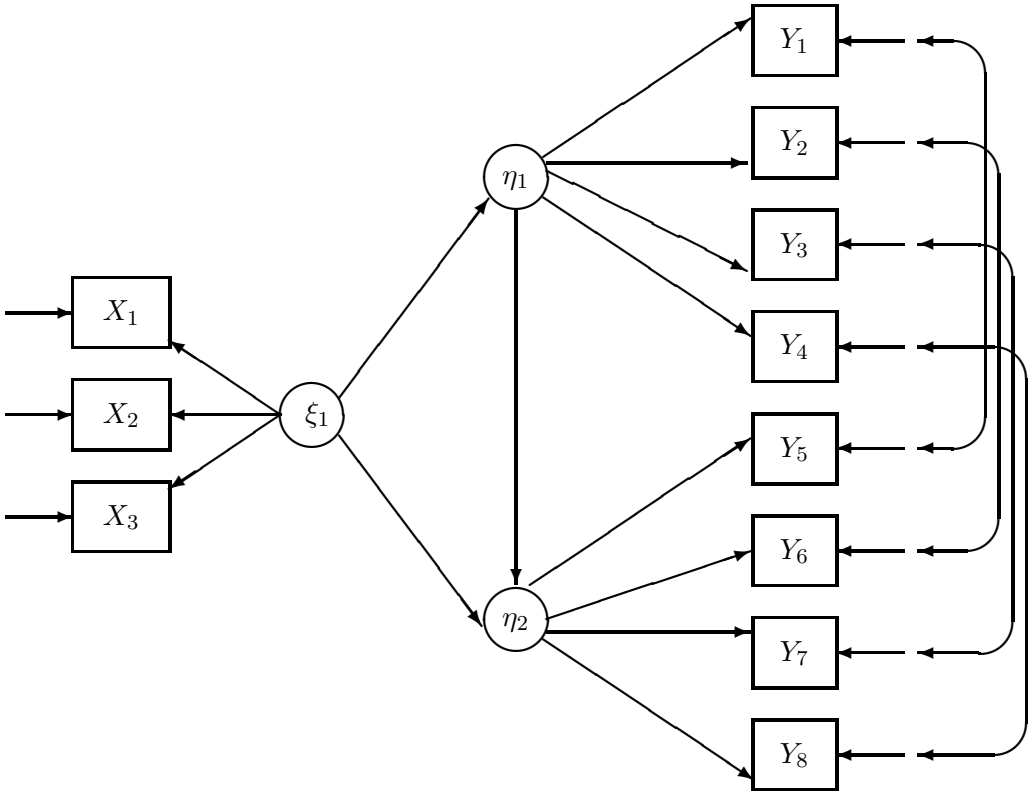

Figure 2: Panel Model of Democracy and Industrialization

- η_1 Democracy in 1960 (Latent variable Dem60)
- η_2 Democracy in 1965 (Latent variable Dem65)
- ξ Level of industrialization in 1960 (Latent variable Indus)

$y_1 - y_4$ are taken as indicators of the latent variable Dem60 (Democracy 1960) and $y_5 - y_8$ are taken as indicators of the latent variable Dem65 (Democracy 1965). $x_1 - x_3$ are taken as indicators of the latent variable Indus (Industrialization 1960). Data on $y_1 - x_3$ are available for 75 developing countries. These data are in the file **POLIDEM.PSF**.

2.1 Latent Variables Scores

The following SIMPLIS syntax file (**BA1a.SPL**) will estimate scores on Demo60, Demo65, and Indus for each country in file **POLIDEMnew.PSF**. Slightly different versions of **BA1a.SPL**, are used in sections 2.2, 2.3, and 2.4, namely **BA1b.SPL**, **BA1c.SPL**, and **BA1d.SPL**.

```
Industrialization-Democracy Example
Raw Data from file POLIDEM.PSF
Latent Variables: Dem60 Dem65 Indus
Relationships:
Y1= 1*Dem60
Y2-Y4 = Dem60
Y5 = 1*Dem65
Y6-Y8 = Dem65
X9 = 1*Indus
X10-X11 = Indus
Dem60  = Indus
Dem65 = Dem60 Indus
Set Dem60 -> Y2 = Dem65 -> Y6
Set Dem60 -> Y3 = Dem65 -> Y7
Set Dem60 -> Y4 = Dem65 -> Y8
Let the errors of Y5 and Y1 be correlated
let the errors of Y6 and Y2 be correlated
Let the errors of Y7 and Y3 be correlated
Let the errors of Y8 and Y4 be correlated
PSFfile POLIDEM.PSF
Path Diagram
End of Problem
```

The variables y_1–y_4 are the same variables as y_5–y_8 measured at two points in time. So Bollen & Arminger (1991) assume that their loadings on η_1 and η_2 are the same. This is specified by the lines

```
Set Dem60 -> Y2 = Dem65 -> Y6
Set Dem60 -> Y3 = Dem65 -> Y7
Set Dem60 -> Y4 = Dem65 -> Y8
```

The loadings of y_1 and y_5 are set to 1. Furthermore, they assume that the measurement errors of corresponding y-variables are correlated. This is specified by the lines

```
Let the errors of Y5 and Y1 be correlated
let the errors of Y6 and Y2 be correlated
Let the errors of Y7 and Y3 be correlated
Let the errors of Y8 and Y4 be correlated
```

After this run is completed the file **POLIDEMnew.PSF** contains the following variables

```
Y1 Y2 Y3 Y4 Y5 Y6 Y7 Y8 X1 X2 X3 Dem60 Dem65 Indus
```

It would be useful to know which country each row belongs to. We do not have this information. However, one can construct a variable COUNTRY which runs from 1 to 75[2] This is done with the following PRELIS syntax file (**BA2.PR2**) which at the same time constructs scores on another latent variable Diff = Dem65 - Dem60.

[2]The variable TIME is always available in PRELIS. It assigns values $1, 2, \ldots, N$ to the cases. It is intended mainly for time series. Hence its name TIME.

```
SY=POLIDEMnew.PSF
New COUNTRY=TIME
New Diff=Dem65-Dem60
CO ALL
Select COUNTRY Dem60 Dem65 Diff
OU RA=DEMDIFF.PSF
```

The file **DEMDIFF.PSF** contains the following variables

```
COUNTRY Dem60 Dem65 Diff
```

With this file one can do various things to find out which countries have most democracy or which countries increased or decreased their democracy between 1960 and 1965. For example, do a bivariate line plot of `Diff` against `COUNTRY`. This shows that country 2 increased democracy most and country 30 had the largest decrease. This can also be seen by sorting Diff in descending order. This shows that country 2 has a Diff value of 1.69 and country 30 has a Diff value of -1.51. These are the best and worst countries. The second best and second worst countries are the countries 22 and 34. These have Diff values of 1.39 and -1.38, respectively.

2.2 Observational Residuals

The observational residuals can be obtained by adding the line

```
Estimate Residuals
```

in the file **BA1a.SPL**, see file **BA1b.SPL**. Running this produces a new file **POLIDEM-new.PSF** containing the following variables (Note that the previous file**POLIDEMnew.PSF** will be overwritten)

```
Y1 Y2 Y3 Y4 Y5 Y6 Y7 Y8 X1 X2 X3 Dem60 Dem65 Indus
R_Y1 R_Y2 R_Y3 R_Y4 R_Y5 R_Y6 R_Y7 R_Y8 R_X1 R_X2 R_X3 R_Dem60 R_Dem65
```

For example, `R_Y5` is the estimate of the measurement error ϵ_5, `R_X3` is the estimate of the measurement error δ_3, and `R_Dem64` is the estimate of the structural error ζ_2 in the LISREL model. One can use these variables in much the same way as described in the previous section. For example, one can add the country code and plot any of the observational residuals against `COUNTRY` or sort any of these residuals in ascending or descending values. This can be done for the purpose of finding errors in the data or other outliers or to determine countries with largest or smallest residuals.

2.3 Standardized Latent Variable Scores

One can also obtain estimates of the standardized latent variables. This is easy to do. Just delete the `1*` in three places in **BA1a.SPL** (see file **BA1c.SPL**) and rerun that file. This gives a solution in which the three latent variables Dem60, Dem65, and Indus have unit variances. The correlation matrix of the latent variables is given in the output file **BA1c.OUT** as

```
         Covariance Matrix of Latent Variables

              Dem60      Dem65      Indus
            --------   --------   --------
    Dem60       1.00
    Dem65       0.94       1.00
    Indus       0.45       0.56       1.00
```

Furthermore, the resulting file **POLIDEMnew.PSF** now contains estimates of the standardized latent variables. Next run the following PRELIS syntax file (**BA3.PR2**):

```
SY=POLIDEMnew.PSF
Select Dem60 Dem65 Indus
OU MA=KM XU
```

This will compute the correlation matrix of the latent variable scores. The output file **BA3.OUT** verifies that this correlation matrix is indeed equal to the correlation matrix of the latent variables in the model as stated in the introduction:

```
 Correlation Matrix

              Dem60      Dem65      Indus
            --------   --------   --------
    Dem60       1.000
    Dem65       0.945      1.000
    Indus       0.449      0.560      1.000
```

PRELIS uses three decimals in the output whereas LISREL (SIMPLIS) uses two decimals by default. One can change that by putting the line

```
Number of decimals: 3
```

in the SIMPLIS command file.

2.4 Standardized Observational Residuals

The latent variable scores and observational residuals considered in sections 2.1 and 2.2 depend on the unit of measurement in the observed y- and x-variables. This is useful if these units have some definite meaning. However, sometimes these units are the result of rather arbitrary scaling of the observed variables, in which case it may be useful to standardize the observational residuals in some way. This is discussed by Bollen & Arminger (1991) who consider different ways of standardizing them.

One way to standardize the observational residuals is to rescale them such that they have zero means and unit standard deviations in the sample. For this purpose we have added a command to standardize variables in PRELIS:

```
SV varlist
```

will standardize the variables in `varlist`. One can then save the standadized variables in a separate data file. For example, to standardize the observational residuals in the file **POLIDEMnew.PSF** obtained by running **BA1b.SPL** and save them in the file **POLIDEMstdres.PSF**, use the following PRELIS command file

```
SY=POLIDEMnew.PSF
SE R_Y1-R_Dem65
SV ALL
OU RA=POLIDEMstdres.PSF XU
```

Another standardization is to standardize the residuals by answering the question: What would the estimates of the residuals be if all variables, observed as well as latent, were standardized?

While it is easy to estimate a completely standardized solution in LISREL, using either SIMPLIS or LISREL syntax, see p. 184 in Jöreskog & Sörbom (1999a) or p. 93 in Jöreskog & Sörbom (1999b), it is not quite as easy to obtain estimates of the residuals in a standardized solution. To obtain such standardized residuals, one must have the observed variables standardized in the **PSF** file.

The following PRELIS syntax file

```
SY=POLIDEM.PSF
SV all
OU RA=POLIDEMstd.PSF XU
```

will standardize all the variables in **POLIDEM.PSF** and save the standardized variables in the file **POLIDEMstd.PSF**.

To estimate the standardized observational residuals, one can now run **BA1b.SPL** with **POLIDEM.PSF** replaced by **POLIDEMstd.PSF** in two places, see file **BA1d.SPL**.

References

Anderson, T.W., & Rubin, H. (1956) Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium*, Volume V. Berkeley: University of California Press.

Bartholomew, D., & Knott, M. (1999) *Latent Variable Models and Factor Analysis*. London: Arnold.

Bartlett, M.S. (1937) The statistical conception of mental factors. *British Journal of Psochology*, **28**,97–104.

Bollen, K.A. (1989) *Structural Equations with Latent Variables*. Wiley.

Bollen, K.A. & Arminger, G. (1991) Observational residuals in factor analysis and structural equation models. *Sociological Methodology*, **21**, 235–262.

Jöreskog K. G. (2000) Latent variable scores.
Available at `http://www.ssicentral.com/lisrel/advancedtopics.html`.

Jöreskog, K.G. & Sörbom, D. (1999a) LISREL 8*: Structural Equation Modeling with the SIMPLIS Command Language*. Lincolnwood, IL: Scientific Software International.

Jöreskog, K.G. & Sörbom, D. (1999b) LISREL 8 *User's Reference Guide*. Lincolnwood, IL: Scientific Software International.

Jöreskog, K.G., Sörbom, D., Du Toit, S., & Du Toit M. (2003) LISREL 8*: New Statistical Features*. Third printing with revisions. Lincolnwood, IL: Scientific Software International.

Jöreskog, K.G., & Yang, F. (1996). Nonlinear structural equation models: The Kenny-Judd model with interaction effects. Pp. 57-88 in G.A. Marcoulides & R.E. Schumacker (Eds): *Advanced structural equation modeling: Issues and techniques.* Lawrence Erlbaum Associates, Publishers.

Kenny, D.A., & Judd, C.M. (1984) Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, 96, 201-210.

Lawley, D.N., & Maxwell, A.E. (1971) *Factor Analysis as a Statistical Method*, (2nd edition). London: Butterworths.

Marsh, H.W., Wen, Z., & Hau, K.T. (2004) Structural equation models of latent interactions: Evaluation of alternative strategies and indicator construction. *Psychological Methods*, **9**, 275–300.

Thomson, G.H. (1939) *The Factorial Nature of Human Ability*. New York: Houghton-Mifflin.