LISREL with incomplete data

Incomplete data is a common problem in social science investigations. In estimating models, researchers often need to combine data from two or more samples or subsamples, each with a somewhat different set of variables. LISREL multi sample option may be used to deal with incomplete data problems. The following interesting example was provided by Allison (1987). (Source: Bielby, et al. (1977), Allison (1987)).

"*Suppose the aim is to estimate the correlation between father's occupational status (FAOC) and father's education attainment (FAEF) for black men in the U.S. Using a sample of 2020, Bielby et al., (1977) estimated that correlation to be 0.433. They recognized, however, that this correlation be attenuated by random measurement error. To estimate and possibly correct for this error, they took a random subsample of 348 from the original sample of 2020 black males and reinterviewed them approximately three weeks later. Consequently, their original sample can be divided into two groups: a small subsample of 348 with complete data and a larger subsample 1672 with incomplete data*" (Allison, 1987, p.84).

Therefore, the complete data sample has two indicators x_1 and x_2 of FAOC and two indicators x_3 and x_4 of FAED, whereas the incomplete data sample has only data on x_1 and x_3. The design of the study suggests that the missing data are missing at random.

Table: Means and covariance matrices for measures of father's occupation and education

	Complete data subsample				Incomplete data subsample			
	Father's occupation		Father's education		Father's occupation		Father's education	
	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
x_1	180.90				217.27			
x_2	1226.77	217.56			0.00	1.00		
x_3	23.96	30.20	16.24		25.57	0.00	16.16	
x_4	22.86	30.47	14.36	15.13	0.00	0.00	0.00	1.00
Mean	16.52	17.39	6.65	6.75	16.98	0.00	6.83	0.00

Data on x_2 and x_4 are missing in the incomplete data sample

The table above gives sample variances, covariances, and means for the two groups. For the missing variables, pseudo values of 1 have been entered for the variances and pseudo-values of 0 for the covariances and the means. The use of such pseudo-values was suggested originally by Jöreskog (1971a). Allison (1987) showed that, when combined with pseudo-

values for appropriate parameters, such pseudo-values still produce correct ML estimates in LISREL. Allison analyzed moment matrices as described in the next example. Here we use the extended LISREL model, which makes it much easier. The model is a Submodel 1 with parameters $\boldsymbol{\tau}_x$, $\boldsymbol{\Lambda}_x$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}_\delta$:

$$\boldsymbol{\tau}_x = \begin{bmatrix} * \\ * \\ * \\ * \end{bmatrix} \quad \boldsymbol{\Lambda}_x = \begin{bmatrix} 1 & 0 \\ * & 0 \\ 0 & 1 \\ 0 & * \end{bmatrix} \quad \boldsymbol{\Phi} = \begin{bmatrix} * & \\ * & * \end{bmatrix} \quad \boldsymbol{\Theta}_g = \begin{bmatrix} * \\ * \\ * \\ * \end{bmatrix}.$$

The two latent variables ξ_1 and ξ_2 represent true FAOC and FAED, respectively, the two 1's in $\boldsymbol{\Lambda}_x$ define the scales for these. For the incomplete sample, we set $\tau_2^{(x)}$, $\tau_4^{(x)}$, $\lambda_{21}^{(x)}$ and $\lambda_{42}^{(x)}$ equal to 0 and $\theta_2^{(\delta)}$ and $\theta_4^{(\delta)}$ equal to 1. All free parameters (the *'s in the model above) are constrained to be equal across subsamples.

The command file (**EX103.LIS** in the **LISREL Example** folder) is:

```
     FATHER SES   COMPLETE DATA
DA NG=2 NI=4 NO=348
CM FI=EX103.DAT
ME FI=EX103.DAT
MO NX=4 NK=2 TX=FR
VA 1 LX 1 1 LX 3 2
FR LX 2 1 LX 4 2
MA PH
100 25 10
MA TX
16 16 6 6
OU SS
FATHERS SES   INCOMPLETE DATA
DA NO=1672
CM FI=EX103.DAT
ME FI=EX103.DAT
MO PH=IN
FI TX 2 TX 4 TD 2 TD 4
VA 1 LX 1 1 LX 3 2 TD 2 TD 4
EQ TX 1 1 TX 1
EQ TX 1 3 TX 3
EQ TD 1 1 1 TD 1
EQ TD 1 3 3 TD 3
MA TX
16 0 6 0
OU DF=-9
```

As this is a somewhat unusual model, LISREL has problems generating good starting values. Starting values are therefore given for $\boldsymbol{\tau}^{(x)}$ and $\boldsymbol{\Phi}$, using MA commands.

The results agree almost exactly with those reported by Allison (1987). The goodness-of-fit is 7.74 with 15 degrees of freedom. However, the degrees of freedom are incorrect, because LISREL has counted the nine pseudo-values in the data as real data values. The correct degrees of freedom should be six. The value of χ^2 still represents a very good fit.

The estimated covariance between true FAOC and true FAED is 25.18 with a standard error of 1.41. This should be compared with the estimate 23.21 with a standard error of 3.13 obtained from the complete data sample alone. By using the incomplete

data sample, the standard error is reduced to less than half. Thus, there is a major gain in precision by using all available data.

We may also compare estimated correlations between true FAOC and true FAED. Based on both samples, this is 0.62. This may be compared with the attenuated estimate of 0.43 reported by Bielby et al. (1977)