Generating non-normal variables

The following example generates six variables $x_1, x_2, ..., x_6$ having the same population covariance matrix as in the previous example. These six variables are generated from $v_1, v_2, ..., v_6$, which are independent where

- v_1 is normal with mean 0 and variance 1
- v_2 is χ^2 with 3 degrees of freedom
- v_3 is a three-point distribution with probabilities ½, 1/3, 1/6, at 0, 1, and 2, respectively.
- v_4 is a uniform discrete distribution with probability 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 μ and σ are the mean and standard deviation of v, and v^* replaces v. Finally, they are transformed by the same matrix $\mathbf{T}$ as before (see generating normal variables example). The input file **SIMEX21.PRL** from the **PRELIS Examples** folder 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 $x_1, x_2, ..., x_6$ is not ideal because they are linear combinations of $v_1, v_2, ..., v_6$ 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 example.

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 $y_1, y_2, ..., y_p, x_1, x_2, ..., x_q$ can be generated as follows:

Step 1 Generate values on $\xi_1, ..., \xi_n$ and $\varsigma_1, ..., \varsigma_m$.

Step 2 Generate values on $\varepsilon_1, ..., \varepsilon_p$.

Step 3 Generate values on $\delta_1, ..., \delta_p$.

Step 4 Generate values on $\eta_1, ..., \eta_m$ from the structural equations in the LISREL model.

Step 5 Generate values on $y_1, ..., y_p$ from the measurement model for the y-variables in the LISREL model.

Step 6 Generate values on $x_1, ..., x_p$ from the measurement model for the x-variables in the LISREL model.

For notation and formulas, see Jöreskog & Sörbom (1989). 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.

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 $\begin{bmatrix} \xi_1 & \xi_2 & \xi_3 \end{bmatrix}$ is

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

and the covariance matrix of $\begin{bmatrix} \delta_1 & \delta_2 & \delta_3 & \delta_4 & \delta_5 & \delta_6 \end{bmatrix}$ is

$$\mathbf{\Theta}_\delta = diag\begin{bmatrix} \theta_1 & \theta_2 & \theta_3 & \theta_4 & \theta_5 & \theta_6 \end{bmatrix}.$$

We assume that

- ξ_1, ξ_2 and ξ_3 are trivariate normal with zero means
- δ_i is uniformly distributed over the interval $(0, \theta_i)$, where $i = 1, ..., 6$
- δ_i is independent of δ_j for $i \neq j$
- δ_i is independent of ξ_j for $i = 1, ..., 6$ and $j = 1, 2, 3$

and the parameter values are

$$(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7) = (0.2, 0.4, 0.7, 0.7, 0.8, 0.2, 0.3),$$
$$(\phi_{11}, \phi_{21}, \phi_{22}, \phi_{33}) = (0.64, 0.5, 0.65, 0.81),$$
$$(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) = (2, 2.5, 3, 3.5, 4, 4.5)$$

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 ξ and δ be non-normal
- one can let some ξ-variable correlate with some δ-variable
- one can let a ξ-variable and a δ-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.