Measurement error in regression models

The full LISREL model combines the measurement model with the model designed to estimate "causal" relationships among directly observed explanatory variables and dependent variables. The full model permits these relationships to be studied when both types of variables are subject to measurement error.

The full model is a combination of a structural equation system among *latent* variables η's and ξ's,

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta},$$

and measurement models for observed *y*'s and *x*'s,

$$\mathbf{y} = \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon}$$
$$\mathbf{x} = \boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta},$$

where all variables, observed and latent, are assumed measured in deviations from their means.

The full model therefore involves the following variables:

Observed Variables:	$\mathbf{y}' = (y_1, y_2, ..., y_p)$	$\mathbf{x}' = (x_1, x_2, ..., y_q)$
Latent Variables:	$\boldsymbol{\eta}' = (\eta_1, \eta_2, ..., \eta_m)$	$\boldsymbol{\xi}' = (\xi_1, \xi_2, ..., \xi_n)$
Error Variables:	$\boldsymbol{\varepsilon}' = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_p)$	$\boldsymbol{\delta}' = (\delta_1, \delta_2, ..., \delta_q)$
	$\boldsymbol{\zeta}' = (\zeta_1, \zeta_2, ..., \zeta_m)$	

The ε's and δ's are called *errors in variables* or measurement errors, and the ζ's are called *errors in equations* or structural disturbance terms.

In addition to the four matrices $\boldsymbol{\Lambda}_y, \boldsymbol{\Lambda}_x, \mathbf{B}$ and $\boldsymbol{\Gamma}$, the model involves the four covariance matrices $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, $\boldsymbol{\Theta}_\varepsilon$ and $\boldsymbol{\Theta}_\delta$, the covariance matrices of $\boldsymbol{\xi}$, $\boldsymbol{\zeta}$, $\boldsymbol{\varepsilon}$, and $\boldsymbol{\delta}$, respectively.

Simple regression:

In the regression of *y* on *x*,

$$y = \gamma_{y.x} x + z \ ,$$

suppose *x* is measured with error:

$$x = \xi + \delta \ ,$$

where δ is the measurement error and ξ is the true value.

Then the relationship between *y* and ξ is

$$y = \gamma\xi + \zeta .$$

Note that the $\gamma_{y.x}$ in the first equation is a *regression parameter*, whereas the γ in the last equation is a *structural parameter*.

If the $\xi, \zeta,$ and δ are mutually uncorrelated, the covariance matrix $\mathbf{\Sigma}$ of (*y*, *x*) is

$$\mathbf{\Sigma} = \begin{bmatrix} \gamma_2\phi + \psi & \\ \gamma\phi & \phi + \theta \end{bmatrix}$$

where $\phi = Var(\xi)$, $\psi = Var(\zeta)$ and $\theta = Var(\delta)$.

It is possible to estimate γ from the elements of the corresponding sample covariance matrix if the relative sizes of the error variances, ϕ and θ, are known or can be estimated. The regression coefficient $\gamma_{y.x}$ is given by

$$\gamma_{y.x} = \frac{Cov(y,x)}{Var(x)} = \frac{\gamma\phi}{\phi + \theta} = \gamma \frac{\phi}{\phi + \theta} = \gamma\rho_{xx}$$

where ρ_{xx} is the reliability of *x*.

It is apparent that the regression parameter $\gamma_{y.x}$ is not equal to the structural parameter, γ, if $\theta > 0$. The sample regression coefficient $\gamma_{y.x}$ is therefore *not a consistent estimator* of γ.

If ρ_{xx} of *x* is known, the structural parameter may be estimated by

$$\hat{\gamma} = \hat{\gamma}_{y.x} / r_{xx}$$

can be used to estimate γ. This reduces the bias in $\hat{\gamma}_{y.x}$ at the expense of an increased sampling variance.

If the reliability of the measure is unknown, it can be estimated by administering two similar measures to the same respondents. The following example illustrates the calculations in the case of a known and an unknown reliability coefficient.

The sample covariance matrix (Härnqvist, 1962) computed from data on a 40-item similarities test for 262 boys who were tested first in grade (x) and later in grade 5 (y) is:

$$S = \begin{array}{c} \begin{matrix} y & \quad x \end{matrix} \\ \begin{bmatrix} 46.886 & \\ 45.889 & 59.890 \end{bmatrix} \end{array}.$$

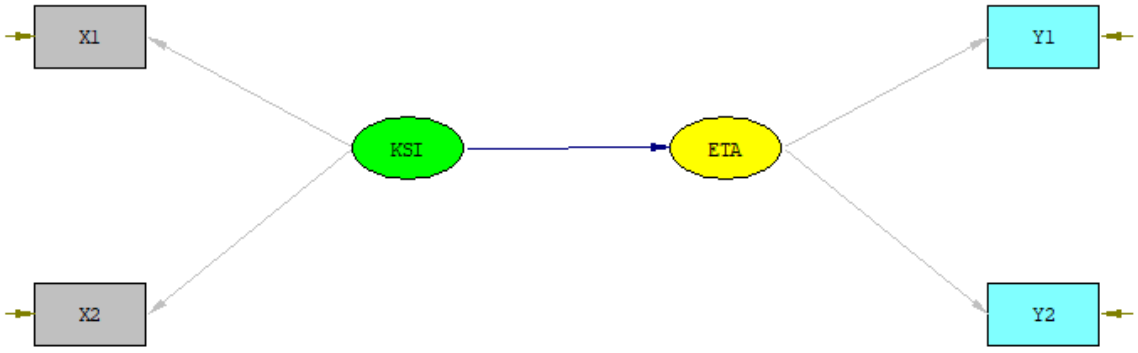
The reliability of the test is given as

$$\hat{\gamma}_{y.x} = \frac{45.889}{59/890} = 0.766.$$

Corrected for attenuation, this becomes

$$\hat{\gamma} = \frac{0.766}{0.896} = 0.855.$$

One way to estimate ρ_{xx} is to split the test into two random 20-item tests and use the score to estimate γ directly using the LISREL model shown below.



Let x_1 and x_2 be random split-halves 20-items tests and use the score to estimate γ directly using the LISREL model shown in the figure above.

$$\mathbf{S} = \begin{array}{c} \begin{matrix} y_1 & y_2 & x_1 & x_2 \end{matrix} \\ \begin{bmatrix} 12.522 & & & \\ 10.405 & 13.554 & & \\ 11.723 & 11.494 & 16.684 & \\ 10.988 & 11.684 & 13.560 & 16.086 \end{bmatrix} \end{array}.$$

The two split-halves are treated as parallel measures. This LISREL command file (**EX51.LIS** in **the LISREL Examples** folder) is:

```
Verbal Ability in Grades 4 and 5
DA NI=4 NO=262
CM FI=EX51.COV
MO NY=2 NX=2 NE=1 NK=1
VA 1 LY(1) LY(2) LX(1) LX(2)
EQ TD(1) TD(2)
EQ TE(1) TE(2)
OU SE
```

The ML estimate of γ obtained by LISREL is

$$\hat{\gamma}_{ML} = 0.846(0.032).$$

Assuming that this contains no bias, we can summarize the relative advantages and disadvantages of the three estimates as in the table below. The ordinary regression estimate $\hat{\gamma}_{y.x}$ has the largest bias and the smallest variance. the ML estimate from LISREL has no bias but a larger variance. Because the reliability of *x*, 0.896, is quite large and because the corrected estimate $\hat{\gamma}$ has a small bias, the MSE of this estimate is comparable to the MSE of the ML estimator in this example.

Multiple regression:

The above principle can be extended to the case of multiple regression with several explanatory variables $x_1, x_2, \ldots, x_n$. Let the regression relationship be

$$y = \gamma_1\xi_1 + \gamma_2\xi_2 + \cdots \gamma_n\xi_n + \zeta.$$

The estimator of γ corresponding to this is

$$\hat{\boldsymbol{\gamma}} = \left(\mathbf{S}_{xx} - \hat{\boldsymbol{\Theta}}_{\delta} \right)^{-1} \mathbf{s}_{y.x},$$

where $\mathbf{S}_{xx}$ is the sample covariance matrix of the x's, $\hat{\boldsymbol{\Theta}}_{\delta}$ is a diagonal matrix of estimated error variances in the x's, and $\mathbf{s}_{y.x}$ is a vector of sample covariances between y and the x's.

In practice, this estimator may fail because $\mathbf{S}_{xx} - \hat{\boldsymbol{\Theta}}_{\delta}$ is often not positive definite. A better approach is to read $\hat{\boldsymbol{\Theta}}_{\delta}$ into LISREL as fixed quantities and let LISREL estimate $\boldsymbol{\gamma}$ from the information provided in the covariance matrix and $\hat{\boldsymbol{\Theta}}_{\delta}$. This approach is illustrated in the following example.

In a study (Warren, White, & Fuller (1974)), of a random sample of 98 managers of farmer cooperatives in Iowa, their role performance, as measured by a role behavior scale (ROLBEHAV) was assumed to be linearly related to four variables:

x_1: Knowledge of economic phases of management directed toward profit making in a business and product knowledge (KNOWLEDG)

x_2: Value Orientation: tendency to rationally evaluate means to an economic end (VALORIEN)

x_3: Role Satisfaction: gratification obtained by the manager from performing the managerial role (ROLSATIS)

x_4: Past Training: amount of formal education (TRAINING)

The covariance matrix of the five variables is given in the table below.

Table: Covariance matrix

	y	x_1	x_2	x_3	x_4
ROLBEHAV	0.0209				
KNOWLEDG	0.0177	0.0520			
VALORIEN	0.0245	0.0280	0.1212		
ROLSATIS	0.0046	0.0044	-0.0063	0.0901	
TRAINING	0.0187	0.0192	0.0353	-0.0066	0.0946

The ordinary least squares (OLS) regression estimates are, with standard errors below,

$$\hat{\boldsymbol{\gamma}}_{\mathbf{y.x}} = 0.230 \quad 0.120 \quad 0.056 \quad 0.110$$
$$(0.053 \quad 0.036 \quad 0.037 \quad 0.039)$$

For the reliabilities of the x-variables given as

$$0.60, \quad 0.64, \quad 0.81, \quad 1,00$$

it is possible to re-estimate $\boldsymbol{\gamma}$ to reduce or eliminate the effects of measurement errors by using a LISREL model of the form

$$\mathbf{x} = \boldsymbol{\xi} + \boldsymbol{\delta} \quad y = \boldsymbol{\gamma}'\boldsymbol{\xi} + \zeta,$$

i.e., we take $\boldsymbol{\Lambda}_y(1\times 1) = 1$, $\boldsymbol{\Lambda}_x(4\times 4) = \mathbf{I}$, $\mathbf{B}(1\times 1) = 0$, $\boldsymbol{\Gamma}(1\times 4) = \boldsymbol{\gamma}'$, $\boldsymbol{\Phi} = Cov(\boldsymbol{\xi})$, unconstrained, $\boldsymbol{\Psi}(1\times 1) = \text{var}(\zeta)$, $\boldsymbol{\Theta}_\varepsilon(1\times 1) = 0$ and $\boldsymbol{\Theta}_\delta$ is diagonal with fixed values 0.0208, 0.0436, 0.0171, 0.0000. These values are obtained by taking 1 minus the reliabilities above times the observed variance.

The LISREL command file for this analysis is (**EX52A.LIS**):

```
Role Behavior of Form Managers, Part A
DA NI=5 NO=98
CM FI=EX52A.COV
LA
ROLBEHAV KNOWLEDG VALORIEN ROLSATIS TRAINING
MO NY=1 NE=1 NX=4 NK=4 LY=ID LX=ID TE=ZE TD=FI
MA TD
.0208 .0436 .0171 .0000
MA PH
.0312
.0280  .0776
.0044 -.0063  .0730
.0192  .0353 -.0066  .0946
OU SE AD=OFF
```

The specification AD = OFF on the OU command is necessary because of the fixed zero in TD(4), i.e., because it is assumed that x_4 has no measurement error.

Partial output for this analysis is shown below:

```
      Covariance Matrix of ETA and KSI

            ROLBEHAV   KNOWLEDG   VALORIEN   ROLSATIS   TRAINING
            --------   --------   --------   --------   --------
 ROLBEHAV      0.021
 KNOWLEDG      0.018      0.031
 VALORIEN      0.025      0.028      0.078
 ROLSATIS      0.005      0.004     -0.006      0.073
 TRAINING      0.019      0.019      0.035     -0.007      0.095
```

The disattenuated estimates of γ and standard errors are

$$\hat{\boldsymbol{\gamma}}_{\mathbf{y.x}} = 0.380 \quad 0.152 \quad 0.059 \quad 0.068$$
$$(0.127 \quad 0.079 \quad 0.050 \quad 0.044).$$

Some of these estimates and standard errors differ considerably from the OLS estimates.

If there are two or more *indicators* for each ξ, the measurement errors as well as the structural parameters can be estimated directly from the data. The procedure is illustrated in the following continuation of the previous example.

Because multiple-item scales were used to measure *y*, x_1, x_2, and x_3, responses to the items could be assigned randomly into two parallel halves. The full covariance matrix of these split-halves is given in the table below.

	y_1	y_2	x_{11}	x_{12}	x_{21}	x_{22}	x_{31}	x_{32}	x_4
y_1	0.0271								
y_2	0.0172	0.0222							
x_{11}	0.0219	0.0193	0.0876						
x_{12}	0.0164	0.0130	0.0317	0.0568					
x_{21}	0.0284	0.0294	0.0383	0.0151	0.1826				
x_{22}	0.0217	0.0185	0.0356	0.0230	0.0774	0.1473			
x_{31}	0.0083	0.0011	-0.0001	0.0055	-0.0087	-0.0069	0.1137		
x_{32}	0.0074	0.0015	0.0035	0.0089	-0.0007	-0.0088	0.0722	0.1024	
x_4	0.0180	0.0194	0.0203	0.0182	0.0563	0.0142	-0.0056	-0.0077	0.0946

These values can be used to estimate the true regression equation

$$\eta = \gamma_1\xi_1 + \gamma_2\xi_2 + \gamma_3\xi_3 + \gamma_4\xi_4 + \zeta,$$

using the following measurement models.

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \eta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

$$\begin{bmatrix} x_{11} \\ x_{12} \\ x_{21} \\ x_{22} \\ x_{31} \\ x_{32} \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1.2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \end{bmatrix} + \begin{bmatrix} \delta_{11} \\ \delta_{12} \\ \delta_{21} \\ \delta_{22} \\ \delta_{31} \\ \delta_{32} \\ 0 \end{bmatrix}.$$

The value 1.2 in the last equation reflects the fact that x_{32} has six items whereas x_{31} has only five.

The latent variables are:

η = role behavior
ξ_1 = knowledge
ξ_2 = value orientation
ξ_3 = role satisfaction
ξ_4 = past training

The observed variables are (the number of items in each split-half is given in parentheses):

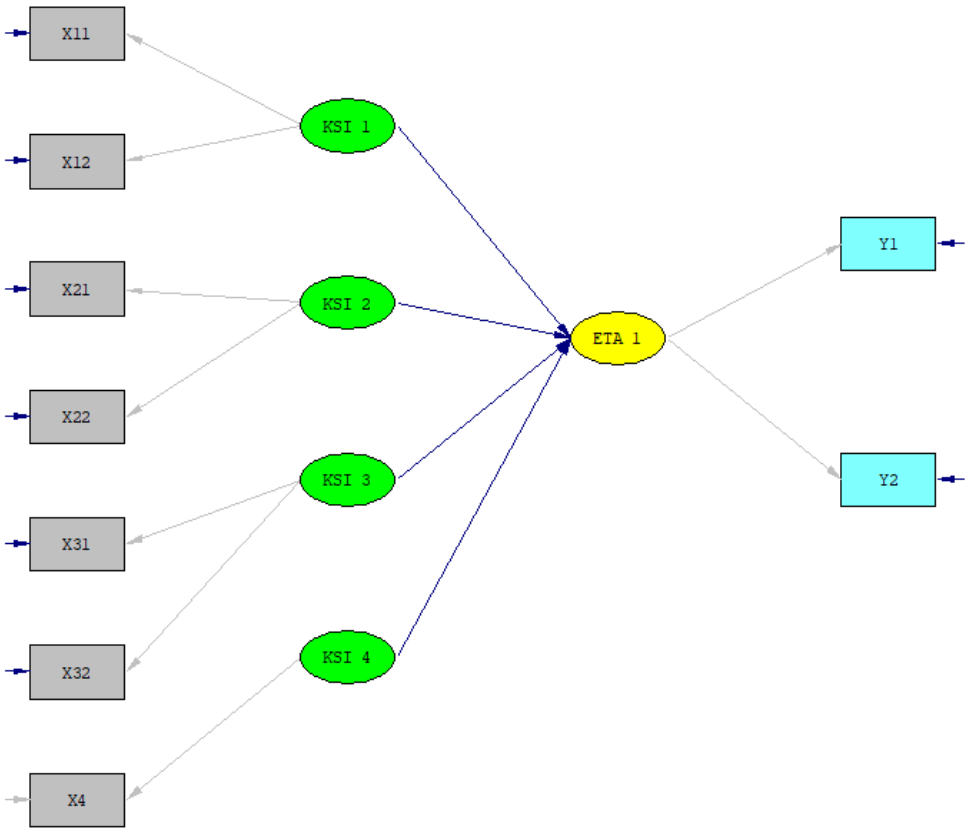
y_1 = a split-half measure of role behavior (12)
y_2 = a split-half measure of role behavior (12)
x_{11} = a split half measure of knowledge (13)
x_{12} = a split half measure of knowledge (13)
x_{21} = a split-half measure of value orientation (15)
x_{22} = a split-half measure of value orientation (15)
x_{31} = a split-half measure of role satisfaction (5)
x_{32} = a split-half measure of role satisfaction (6)
$x_4 = \xi_4$ = a measure of past training.

The LISREL command file (**LISEX52B.LIS**) is:

```
Role Behavior of Form Managers, Part B
DA NI=9 NO=98
CM FI=EX52B.COV
LA
Y1 Y2 X11 X12 X21 X22 X31 X32 X4
```

```
MO NY=2 NE=1 NX=7 NK=4
FI TD 7
VA 1 LY 1 LY 2 LX 1 1 LX 2 1 LX 3 2 LX 4 2 LX 5 3 LX 7 4
VA 1.2 LX 6 3
OU SE AD=OFF
```

The path diagram for this model is shown below.



The fit of the model is χ^2 = 26.97 with 22 degrees of freedom, which represents a rather good fit. The ML estimates of the γ's and their standard errors (below) are

GAMMA

	KSI 1	KSI 2	KSI 3	KSI 4
ETA 1	0.350	0.168	0.045	0.071
	(0.132)	(0.079)	(0.053)	(0.044)
	2.652	2.135	0.848	1.611

When compared to the ordinary least squares (OLS) estimates, previously given for the regression of y on x_1, x_2, x_3, and x_4, considerable bias in the OLS estimates is evident, even though their standard errors are smaller.

Estimates of the true and error score variances for each observed measure are also obtained. The reliability of the measures computed from these estimates are

y	x_1	x_2	x_3
0.82	0.60	0.64	0.81

The model defined in this section can be generalized directly to the case when there are several jointly dependent variables $\boldsymbol{\eta}$. The only differences will be that $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ will be replaced by matrices $\boldsymbol{\Lambda}_y$ and $\boldsymbol{\Gamma}$, respectively, and ψ by a full symmetric positive-definite matrix $\boldsymbol{\Psi}$.