Missing Data Examples

Mplus

(some output omitted to save space)

```
INPUT INSTRUCTIONS

    title:  CFA of three negative exchanges factors;

    data:  file=missing.dat; format=free;

    variable:  names = neg6 neg26 neg30 neg35
        neg11 neg12 neg13 neg14
        neg16 neg17 neg19 neg20;

        missing=neg6-neg20 (-99);

    usevariables=neg6 neg26 neg30 neg35
        neg11 neg12 neg13 neg14
        neg16 neg17 neg19 neg20;

   analysis: type=general;

  !Note: by default in Mplus (version 5 and later), when missing data are present,
  !FIML estimation is used. If the data are non-normal (as they appear to
  ! be in this case), a robust estimation approach should be used (Yuan & Bentler, 2000).
  !  Specify this by adding ESTIMATOR=MLR to the analysis line.

   model: hostile by neg6-neg35;
          badadv by neg11-neg14;
         demands by neg16-neg20;

   output: stdyx patterns;
  !including patterns on the output line provides
  ! descriptive statistics about the missing data patterns;

INPUT READING TERMINATED NORMALLY

CFA of three negative exchanges factors;

SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                         275

Number of dependent variables                                   12
Number of independent variables                                  0
Number of continuous latent variables                            3

Estimator                                                       ML
Information matrix                                        OBSERVED
Maximum number of iterations                                 1000
Convergence criterion                                   0.500D-04
Maximum number of steepest descent iterations                  20
Maximum number of iterations for H1                          2000
Convergence criterion for H1                            0.100D-03

SUMMARY OF DATA

     Number of missing data patterns             7

SUMMARY OF MISSING DATA PATTERNS
     MISSING DATA PATTERNS (x = not missing)
           1  2  3  4  5  6  7
 NEG6      x  x  x  x  x  x
 NEG26     x  x  x  x  x     x
 NEG30     x  x  x  x     x  x
 NEG35     x  x  x  x  x  x  x
 NEG11     x  x  x     x  x
 NEG12     x  x  x  x  x  x
 NEG13     x  x     x  x  x
 NEG14     x  x  x  x  x  x
 NEG16     x  x  x  x  x  x
 NEG17     x     x  x  x
 NEG19     x  x  x  x  x  x  x
 NEG20     x  x  x  x  x  x  x
```

MISSING DATA PATTERN FREQUENCIES

Pattern	Frequency	Pattern	Frequency	Pattern	Frequency
1	194	4	1	7	8
2	69	5	1		
3	1	6	1		

COVARIANCE COVERAGE OF DATA

Minimum covariance coverage value 0.100

PROPORTION OF DATA PRESENT

Covariance Coverage

	NEG6	NEG26	NEG30	NEG35	NEG11
NEG6	0.971				
NEG26	0.967	0.996			
NEG30	0.967	0.993	0.996		
NEG35	0.971	0.996	0.996	1.000	
NEG11	0.967	0.964	0.964	0.967	0.967
NEG12	0.971	0.967	0.967	0.971	0.967
NEG13	0.967	0.964	0.964	0.967	0.964
NEG14	0.971	0.967	0.967	0.971	0.967
NEG16	0.971	0.967	0.967	0.971	0.967
NEG17	0.716	0.716	0.713	0.716	0.713
NEG19	0.971	0.996	0.996	1.000	0.967
NEG20	0.971	0.996	0.996	1.000	0.967

Covariance Coverage

	NEG12	NEG13	NEG14	NEG16	NEG17
NEG12	0.971				
NEG13	0.967	0.967			
NEG14	0.971	0.967	0.971		
NEG16	0.971	0.967	0.971	0.971	
NEG17	0.716	0.713	0.716	0.716	0.716
NEG19	0.971	0.967	0.971	0.971	0.716
NEG20	0.971	0.967	0.971	0.971	0.716

Covariance Coverage

	NEG19	NEG20
NEG19	1.000	
NEG20	1.000	1.000

THE MODEL ESTIMATION TERMINATED NORMALLY

MODEL FIT INFORMATION

Number of Free Parameters 39

Loglikelihood
H0 Value -2884.221
H1 Value -2806.647

Information Criteria
Akaike (AIC) 5846.443
Bayesian (BIC) 5987.497
Sample-Size Adjusted BIC 5863.835
(n* = (n + 2) / 24)

Chi-Square Test of Model Fit
Value 155.149
Degrees of Freedom 51
P-Value 0.0000

RMSEA (Root Mean Square Error Of Approximation)
Estimate 0.086
90 Percent C.I. 0.071 0.102
Probability RMSEA <= .05 0.000

CFI/TLI

```
          CFI                                0.938
          TLI                                0.920

Chi-Square Test of Model Fit for the Baseline Model
          Value                           1758.226
          Degrees of Freedom                   66
          P-Value                          0.0000

SRMR (Standardized Root Mean Square Residual)
          Value                              0.049

MODEL RESULTS
                                                                Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

 HOSTILE  BY
    NEG6               1.000      0.000    999.000    999.000
    NEG26              1.256      0.101     12.383      0.000
    NEG30              1.204      0.101     11.924      0.000
    NEG35              1.015      0.094     10.749      0.000

 BADADV   BY
    NEG11              1.000      0.000    999.000    999.000
    NEG12              1.017      0.089     11.476      0.000
    NEG13              1.375      0.114     12.028      0.000
    NEG14              1.520      0.125     12.139      0.000

 DEMANDS  BY
    NEG16              1.000      0.000    999.000    999.000
    NEG17              1.024      0.095     10.779      0.000
    NEG19              1.019      0.117      8.696      0.000
    NEG20              0.987      0.100      9.847      0.000

 BADADV   WITH
    HOSTILE            0.194      0.028      6.857      0.000

 DEMANDS  WITH
    HOSTILE            0.256      0.036      7.167      0.000
    BADADV             0.208      0.030      6.835      0.000

Variances
    HOSTILE            0.282      0.046      6.192      0.000
    BADADV             0.240      0.038      6.396      0.000
    DEMANDS            0.312      0.050      6.299      0.000


STANDARDIZED MODEL RESULTS
STDYX Standardization
                                                                Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

 HOSTILE  BY
    NEG6               0.695      0.036     19.278      0.000
    NEG26              0.858      0.022     39.590      0.000
    NEG30              0.826      0.024     34.143      0.000
    NEG35              0.720      0.033     21.738      0.000

 BADADV   BY
    NEG11              0.716      0.035     20.477      0.000
    NEG12              0.763      0.031     24.502      0.000
    NEG13              0.813      0.027     30.143      0.000
    NEG14              0.820      0.026     30.965      0.000

 DEMANDS  BY
    NEG16              0.719      0.037     19.666      0.000
    NEG17              0.815      0.033     24.837      0.000
    NEG19              0.616      0.044     13.860      0.000
    NEG20              0.698      0.038     18.216      0.000

 BADADV   WITH
    HOSTILE            0.745      0.038     19.390      0.000

 DEMANDS  WITH
    HOSTILE            0.863      0.031     27.419      0.000
    BADADV             0.758      0.041     18.304      0.000
```

lavaan

(note: some output omitted)

```
> library(lavaan) # always call lavaan library first
> # first time use on the computer, install the lavaan package with the following command
> # install.packages("lavaan", dependencies=TRUE)
>
> ## Missing data estimation example
>
> missdat = read.table("c:/jason/mplus/semclass/missing.dat", header=FALSE)
> names(missdat) = c("neg6","neg26","neg30","neg35",
+       "neg11","neg12","neg13","neg14",
+       "neg16","neg17","neg19","neg20")
> missdat[missdat == -99] <- NA
>
> model = '
+        hostile =~ neg6 + neg26 + neg30 + neg35
+       badadv =~ neg11 + neg12 + neg13 + neg14
+       demands =~ neg16 + neg17 + neg19 + neg20
+ '
> #full information maximum likelihood for normally distributed missing data
> fit = sem(model, data = missdat, missing = "fiml")
> summary(fit,fit.measures=TRUE, rsquare=TRUE, standardized=TRUE)
lavaan (0.5-23.1097) converged normally after  48 iterations

  Number of observations                           275

  Number of missing patterns                         7

  Estimator                                         ML
  Minimum Function Test Statistic              155.149
  Degrees of freedom                                51
  P-value (Chi-square)                           0.000

Model test baseline model:
  Minimum Function Test Statistic             1758.226
  Degrees of freedom                                66
  P-value                                        0.000

User model versus baseline model:
  Comparative Fit Index (CFI)                    0.938
  Tucker-Lewis Index (TLI)                       0.920

Loglikelihood and Information Criteria:
  Loglikelihood user model (H0)              -2884.221
  Loglikelihood unrestricted model (H1)      -2806.647

Root Mean Square Error of Approximation:
  RMSEA                                          0.086
  90 Percent Confidence Interval          0.071  0.102
  P-value RMSEA <= 0.05                          0.000

Standardized Root Mean Square Residual:
  SRMR                                           0.049

Parameter Estimates:
  Information                                 Observed
  Standard Errors                             Standard

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  hostile =~
    neg6              1.000                               0.531    0.695
    neg26             1.256    0.101   12.383    0.000    0.668    0.858
    neg30             1.204    0.101   11.925    0.000    0.640    0.826
    neg35             1.015    0.094   10.749    0.000    0.539    0.720
  badadv =~
    neg11             1.000                               0.490    0.716
    neg12             1.017    0.089   11.476    0.000    0.499    0.763
    neg13             1.375    0.114   12.028    0.000    0.674    0.813
    neg14             1.520    0.125   12.139    0.000    0.745    0.820
  demands =~
    neg16             1.000                               0.559    0.719
    neg17             1.024    0.095   10.779    0.000    0.572    0.815
    neg19             1.019    0.117    8.696    0.000    0.569    0.616
    neg20             0.987    0.100    9.847    0.000    0.551    0.698

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  hostile ~~
    badadv            0.194    0.028    6.857    0.000    0.745    0.745
    demands           0.256    0.036    7.167    0.000    0.863    0.863
  badadv ~~
    demands           0.208    0.030    6.835    0.000    0.758    0.758
```

```
> inspect(fit, 'patterns')
     neg6 neg26 neg30 neg35 neg11 neg12 neg13 neg14 neg16 neg17 neg19 neg20
[1,]    1     1     1     1     1     1     1     1     1     1     1     1
[2,]    1     1     1     1     1     1     1     1     1     0     1     1
[3,]    0     1     1     1     0     0     0     0     0     0     1     1
[4,]    1     0     1     1     1     1     1     1     1     0     1     1
[5,]    1     1     0     1     1     1     1     1     1     1     1     1
[6,]    1     1     1     1     0     1     1     1     1     1     1     1
[7,]    1     1     1     1     1     1     0     1     1     1     1     1
> inspect(fit, 'coverage')
      neg6  neg26 neg30 neg35 neg11 neg12 neg13 neg14 neg16 neg17 neg19 neg20
neg6  0.971
neg26 0.967 0.996
neg30 0.967 0.993 0.996
neg35 0.971 0.996 0.996 1.000
neg11 0.967 0.964 0.964 0.967 0.967
neg12 0.971 0.967 0.967 0.971 0.967 0.971
neg13 0.967 0.964 0.964 0.967 0.964 0.967 0.967
neg14 0.971 0.967 0.967 0.971 0.967 0.971 0.967 0.971
neg16 0.971 0.967 0.967 0.971 0.967 0.971 0.967 0.971 0.971
neg17 0.716 0.716 0.713 0.716 0.713 0.716 0.713 0.716 0.716 0.716
neg19 0.971 0.996 0.996 1.000 0.967 0.971 0.967 0.971 0.971 0.716 1.000
neg20 0.971 0.996 0.996 1.000 0.967 0.971 0.967 0.971 0.971 0.716 1.000 1.000
>
> #Yuan-Bentler robust estimates for nonnormal missing data
> #fit = sem(model, data = missdat, missing = "fiml", estimator="mlr")
> #summary(fit,fit.measures=TRUE, rsquare=TRUE, standardized=TRUE)
```

Missing Values on X Variables

Mplus

When measured exogenous variables (but not indicators for exogenous latent variables) have missing values, the cases with missing data are excluded from the analysis in Mplus. `lavaan` does not exclude cases in this way if you add fixed.x = FALSE to the sem function line.

Below, I test a simple regression model using the data from the path analysis example (see handout “Path Analysis Example: Mplus and lavaan”), which had 196 cases when any cases with missing values were excluded (i.e., listwise deletion). There are 217 cases in the data set overall, however. When a regression model using the path is attempted, a warning indicates that there are 21 cases that are missing data on the x variables (pleased and pc).

```
Mplus VERSION 8.9
MUTHEN & MUTHEN

INPUT INSTRUCTIONS


  title:  Data from social control pilot study (missing on x example);

  data:  file=controlpath2.dat;  format=free;

  variable:  names = pc pleased intent;
             missing=pc pleased intent(-99);

  analysis:  type=general;

  model: intent on pleased pc;

  output:  stdyx patterns;
```

The results produce a warning.

```
*** WARNING
  Data set contains cases with missing on x-variables.
  These cases were not included in the analysis.
  Number of cases with missing on x-variables:  21
   1 WARNING(S) FOUND IN THE INPUT INSTRUCTIONS
```

(The rest of the output is omitted.)

A work around, suggested by Enders (2013) is to specify latent variables with single indicators (loading set to 1 and measurement residual variance set to 0). This model includes the 21 cases missing on the predictors and has a total sample size of 217.

```
INPUT INSTRUCTIONS

  title:  Data from social control pilot study (missing on x example);

  data:  file=controlpath2.dat;  format=free;

  variable:  names = pc pleased intent;
             missing=pc pleased intent(-99);

  analysis:  type=general;

  !Mplus does not include exogenous measured variables
    !  in the analysis when they are missing values;
    !A method of including them is to set up latent variables
    !  for each measured variable, as below;
  model:  !loadings set to 1 and meas res var to 0;
          p by pleased@1;
          pleased@0;
          c by pc@1;
          pc@0;

          intent on p c;

  output:  stdyx patterns;
```

Note, no error messages from missing on X.

```
INPUT READING TERMINATED NORMALLY

Data from social control pilot study (missing on x example);

SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                         217
```

(rest of the output omitted)

R with Missing on X

Error message in R is generated if cases are missing on x.

```
> fit = sem(model, data = missdat, missing = "fiml")
Warning message:
In lav_data_full(data = data, group = group, cluster = cluster,  :
  lavaan WARNING: 21 cases were deleted due to missing values in
                  exogenous variable(s), while fixed.x = TRUE.
```

To include these missing casese, R has a shortcut by adding fixed.x = FALSE.

```
> fit = sem(model, data = missdat, missing = "fiml", fixed.x = FALSE)
```

Specifying Auxiliary Variables

Adding auxiliary variables in either package is relatively simple.

Mplus

```
variable:
    auxiliary = srh1 age (m);
    ! specifies auxiliary variables, (m) is always used
    ! and indicates variables are missing data correlates;
```

Lavaan

```
fitsem <- sem(model13.2a, data=healthmissing, meanstructure=TRUE, missing = "fiml")
fitsemaux <- sem.auxiliary(fitsem, data=healthmissing, aux=c("age", "srh1"), meanstructure=TRUE)
```

References

Enders, C.K. (2013). Analyzing structural equation models with missing data. In G.R. Hancock & R.O. Mueller (Eds.), *Structural equation modeling: A second course, 2nd edition* (pp.493-520). Charlotte, NC: Information Age Publishing.