Continuous variables with missing values

1. Treatment of Missing Values

Missing values and incomplete data are almost unavoidable in in social, behavioral, medical and most other areas of investigation. One can distinguish between three types of incomplete data:

- Unit nonresponse, for example, a person does not respond at all to an item in a questionnaire.
- Subject attrition, for example, when a person falls out of a sample after some time in a longitudinal follow-up study.
- Item nonresponse, for example, a person responds to some but not all items in a questionnaire.

The literature, *e.g.*, Schafer (1997) distinguishes between three mechanisms of nonresponse.

- **MCAR** Missing completely at random
- **MAR** Missing at random
- **MNAR** Missing not at random

Let z_{ij} be any element on the data matrix. Informally, one can define these concepts as

- **MCAR** $\Pr\left(z_{ij} = missing\right)$ does not depend on any variable in the data.
- **MAR** $\Pr\left(z_{ij} = missing\right)$ may depend on other variables in the data but not on z_{ij}. Example: A missing value of a person's income may depend on his/her age and education but not on his/her actual income.
- **MNAR** $\Pr\left(z_{ij} = missing\right)$ depends on z_{ij}. Example: In a questionnaire people with higher income tend not to report their income.

LISREL has several ways of dealing with missing values:

1. Listwise deletion
2. Pairwise deletion
3. Imputation by matching
4. Multiple imputation
    - EM
    - MCMC

5. Full Information Maximum Likelihood (FIML)[1]

Of these methods the first three are ad hoc procedures whereas the last two are based probability models for missingness. As a consequence, the ad hoc methods may lead to biased estimates under MAR and can only be recommended under MCAR.

Listwise deletion means that all cases with missing values are deleted. This leads to a complete data matrix with no missing values which is used to estimate the model. This procedure can lead to a large loss of information in that the resulting sample size is much smaller than the original. Listwise deletion can give biased, inconsistent, and inefficient estimates under MAR. It should only be used under MCAR.

Pairwise deletion means that means and variances are estimated using all available data for each variable and covariances are estimated using all available data for each pair of variables. These means, variances and covariances are then combined to form a mean vector and a covariance matrix which are used to estimate the model. While some efficiency is obtained compared to listwise deletion, it is difficult to specify a sample size N to be used in the estimation of the model, since the variances and covariances are all based on different sample sizes and there is no guaranty that the covariance matrix will be positive definite which is required by the maximum likelihood method. Although pairwise deletion is available in LISREL, it is not recommended. Its best use is for data screening for then it gives the most complete information about the missing values in the data.

Imputation means that real values are substituted for the missing values. Various ad hoc procedures for imputation have been suggested in the literature. One such is imputation by matching which is available in LISREL. It is based on the idea that individuals who have similar values on a set of matching variables may also be similar on a variable with missing values. This will work well if the matching variables are good predictors of the variable with missing values.

Methods 4 and 5 are both based on the assumption of multivariate normality and missingness under MAR. Method 4 uses multiple imputation methods to generate a complete data matrix. The multiple imputation procedure implemented in LISREL is described in detail in Schafer (1997) and uses the EM algorithm and the method of generating random draws from probability distributions via Markov chains (MCMC). The EM algorithm generates one single complete data matrix whereas the MCMC method generates several complete data matrices and uses the average of these. As a consequence, the MCMC method is more reliable than the EM algorithm. In both cases, the complete data matrix can be used to estimate the mean vector and the covariance matrix of the observed variables which can be used to estimate the model. However, in LISREL it is not necessary to do these steps separately as they are done automatically as will be described in what follows.

Method 5 is the default method in LISREL when there are missing data. This is the recommended method for dealing with the problem of missing data. So this is described first.

If the variables have a multivariate normal distribution all subsets of the variables also have that distribution. So the likelihood function for the observed values can be evaluated for each observation without using any missing values.

2. Latent Curve Models: Example of treatment of Prostate Cancer (PSAVAR)

A medical doctor offered all his patients diagnosed with prostate cancer a treatment aimed at reducing the cancer activity in the prostate. The severity of prostate cancer is often assessed by a plasma component known as prostate specific antigen (PSA), an enzyme that is elevated in the presence of prostate cancer. The PSA level was measured regularly every three months. The data contains five repeated measurements of PSA. The age of the patient is also included in the data. Not every patient accepted the offer initially and several patients chose to enter the program

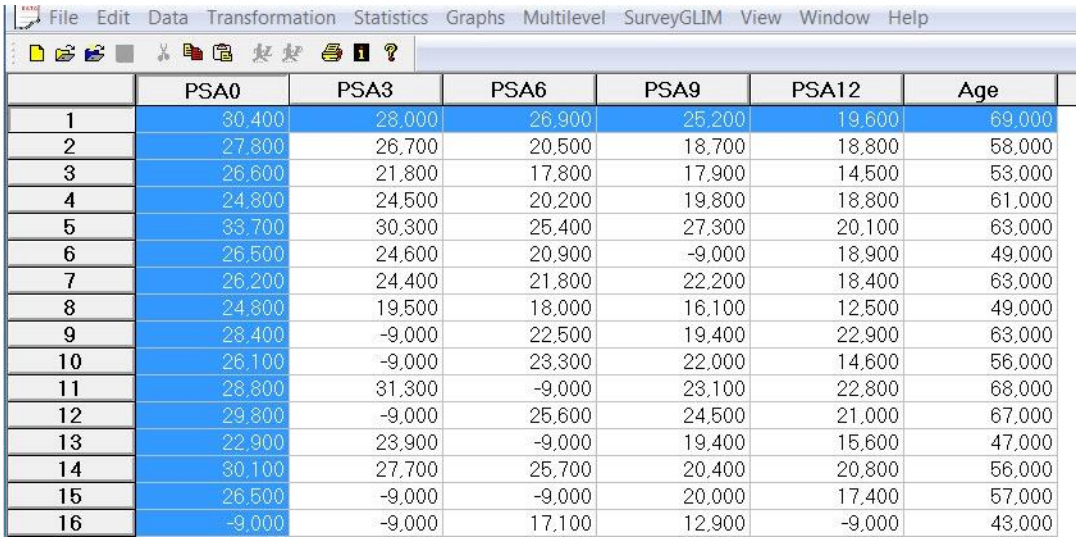
[1] Of course, the maximum likelihood (ML) used earlier is also a full information maximum likelihood method. However, it is convenient to use the term ML for the case of complete data and the term FIML for the case of missing data.

after the first occasion. Some patients, who accepted the initial offer, are absent at some later occasions for various reasons. Thus there are missing values in the data.

The aim of this study is to answer the following questions: What is the average initial PSA value? Do all patients have the same initial PSA value? Is there an overall effect of treatment. Is there a decline of PSA values over time, and, if so, what is the average rate of decline? Do all patients have the same rate of decline? Does the individual initial PSA value and/or the rate of decline depend on the patient's age?

This is a typical example of repeated measurements data, the analysis of which is sometimes done within the framework of multilevel analysis. It represents the simplest type of two-level model but it can also be analyzed as a structural equation model, see Bollen & Curran (2006). In this context it illustates a mean and covariance structure model estimated from longitudinal data with missing values.

The data file for this example is **psavar.lsf**, where missing values are shown as -9.000[2]. This data file is stored in the **SIMPLIS Examples** folder.



	PSA0	PSA3	PSA6	PSA9	PSA12	Age
1	30,400	28,000	26,900	25,200	19,600	69,000
2	27,800	26,700	20,500	18,700	18,800	58,000
3	26,600	21,800	17,800	17,900	14,500	53,000
4	24,800	24,500	20,200	19,800	18,800	61,000
5	33,700	30,300	25,400	27,300	20,100	63,000
6	26,500	24,600	20,900	-9,000	18,900	49,000
7	26,200	24,400	21,800	22,200	18,400	63,000
8	24,800	19,500	18,000	16,100	12,500	49,000
9	28,400	-9,000	22,500	19,400	22,900	63,000
10	26,100	-9,000	23,300	22,000	14,600	56,000
11	28,800	31,300	-9,000	23,100	22,800	68,000
12	29,800	-9,000	25,600	24,500	21,000	67,000
13	22,900	23,900	-9,000	19,400	15,600	47,000
14	30,100	27,700	25,700	20,400	20,800	56,000
15	26,500	-9,000	-9,000	20,000	17,400	57,000
16	-9,000	-9,000	17,100	12,900	-9,000	43,000

In this kind of data it is inevitable that there are missing values. For example, a patient may be on vacation or ill or unable to come to the doctor for any reason at some occasion or a patient may die and therefore will not come to the doctor after a certain occasion. It is seen in that

- Patients 9 and 10 are missing at 3 months
- Patient 15 is missing at 3 and 6 months
- Patient 16 is missing at 0, 3, and 12 months

In the following analysis it is assumed that data are missing at random (MAR), although there may be a small probability that a patient will be missing because his PSA value is high.

Whenever one starts an analysis of a new data set, it is recommended to begin with a data screening. To do so click on **Statistics** at the top of the screen and select **Data Screening** from the **Statistics** menu. This will reveal the following information about the data.

```
Number of Missing Values per Variable
PSA0  PSA3  PSA6  PSA9  PSA12 Age
```

[2] If the data is imported from an external source which already have a missing value code, the missing values will show up in the **lsf** file as -999999.000, which is the global missing data code in LISREL.

```
-------- -------- -------- -------- -------- --------
17    14    13    12    11    0
This table says that there are 17 patients missing initially, 14 missing at 3 months, 13
at 6 months, etc.
Distribution of Missing Values
Total Sample Size =      100
 Number  of  Missing     0      1      2     3
 Values
 Number of Cases         46     43     9     2
```

This table says that there are only 46 patients with complete data on all six occasions. Thus, if one uses listwise deletion 54% of the sample will be lost. 43 patients are missing on one occasion, 9 patients are missing at two occasions, 2 patients are missing on three occasions. This table does not tell on which occasions the patients are missing. The next table gives more complete information about the missing data patterns.

```
Missing Data Map
Frequency PerCent Pattern
46    46.0 0 0 0 0 0 0
9     9.0 1 0 0 0 0 0
8     8.0 0 1 0 0 0 0
2     2.0 1 1 0 0 0 0
8     8.0 0 0 1 0 0 0
2     2.0 1 0 1 0 0 0
2     2.0 0 1 1 0 0 0
9     9.0 0 0 0 1 0 0
1     1.0 1 0 0 1 0 0
1     1.0 1 1 0 1 0 0
1     1.0 0 0 1 1 0 0
9     9.0 0 0 0 0 1 0
1     1.0 1 0 0 0 1 0
1     1.0 1 1 0 0 1 0
```

The columns under Pattern correspond to the variables in the order they are in **psavar.lsf**. A 0 means a non-missing value and a 1 means a missing value. Recall that the last variable is the patient's age. This has no missing values. Here one can see for example that two patients are missing at both 0 and 3 months and another patient is missing at 6 and 9 months.

The following information about the univariate distributions of the variables have been obtained using all available data for each variable, i.e., 83 patients for PSA0, 86 patients for PSA3, etc.

Univariate Summary Statistics for Continuous Variables

```
Variable     Mean St. Dev. Skewness Kurtosis Minimum Freq. Maximum Freq.
--------     ---- --------     ------- -------- ------- ----- ------- -
                                ----
PSA0         31.1645.684       0.068 -0.85219.9001     44.1001
PSA3         30.0366.025       -0.248-0.73214.5001     42.1001
PSA6         27.4436.084       -0.335-0.96113.7001     37.6001
PSA9         25.3336.391       -0.331-1.06610.6001     36.2001
PSA12        23.4066.306       -0.309-1.0699.600 1     35.8001
Age          55.4507.896       -0.329-0.23432.0001     70.0001
```

It is seen that the mean age is 55.45 years and that average initial PSA value is 31.164 with a minimum at 19.9 and maximum at 44.1. At 12 months the corresponding values are 23.406, 9.6, and 35.8, respectively. Thus there is some evidence that the PSA values are decreasing over time.

The model to be estimated is

$$y_{it} = a_i + b_i T_t + e_{it} \quad (1)$$

$$i = 1, 2, ..., N \text{ individuals} \quad (2)$$

$$T_t = \text{Time at occasion } t = 1, 2, ..., n_i \quad (3)$$

$$a_i = \alpha + \gamma_a z_i + u_i \quad (4)$$

$$b_i = \beta + \gamma_b z_i + v_i \quad (5)$$

$$z_i = \text{Covariate observed on individual } i \quad (6)$$

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N\left(\mathbf{0}, \mathbf{\Phi}\right) \quad (7)$$

$$e_{it} \sim N\left(0, \sigma_\varepsilon^2\right) \quad (8)$$

$$y_{it} = (\alpha + \beta T_t + \gamma_a z_i + \gamma_b T_t z_i) + (u_i + v_i T_t + e_{it} \quad (9)$$

An interpretation of this is as follows. Each patient has his own linear growth curve[3], represented by (1) which is the regression of y_{it} on time with intercept a_i and slope b_i varying across patients. In principle, the intercepts a_i and slopes b_i could all be different across patients. It is of interest to know if the intercepts and/or the slopes are equal across patients. The four cases are illustrated in Figure 1. If there is variation in intercepts and/or the slopes across patients, one is interested in whether a covariate z_i (in this case age) can predict the intercept and/or the slope.

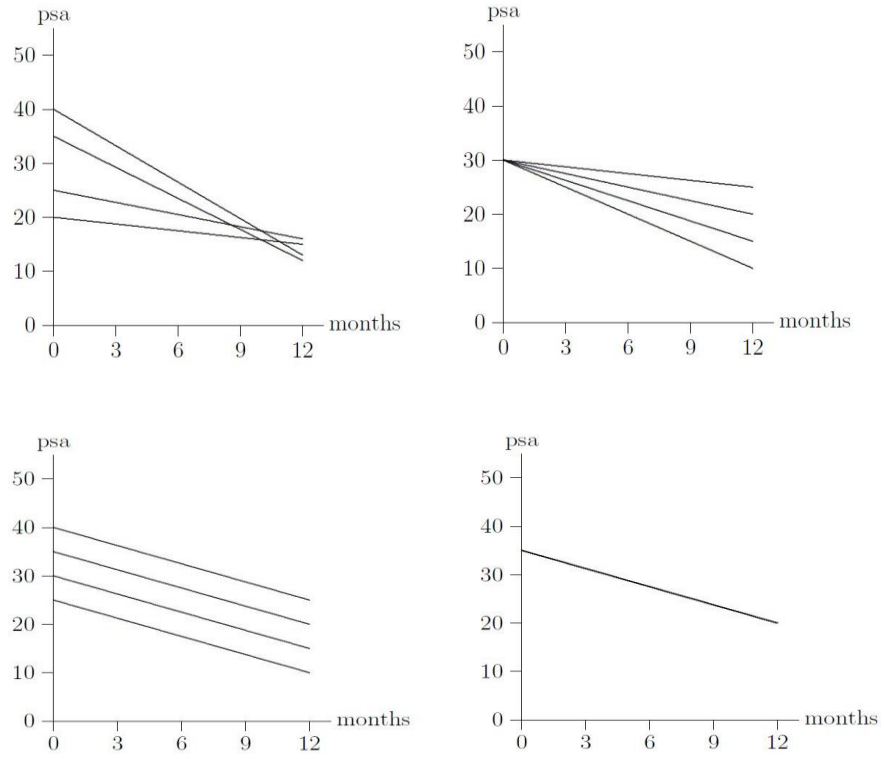


Figure 1: Four Cases of Intercepts and Slopes

[3] In general, the growth curves are not restricted to be linear, but can be quadratic, cubic, or other types of functions of time, see Jöreskog, Sörbom, Du Toit, & Du Toit (2003).

Path diagrams for the models without and with a covariate are illustrated in Figures 2 and 3, respectively, with $T_t = t - 1$ for four occasions.

The model in Figure 1 can be estimated with FIML using the following SIMPLIS syntax file (**psavar1a.spl**):

```
Linear Growth Curve for psavar Data
Raw Data from File psavar.LSF
Latent Variables: a b
Relationships
PSA0 = 1*a 0*b
PSA3 = 1*a 3*b
PSA6 = 1*a 6*b
PSA9 = 1*a 9*b PSA12 = 1*a 12*b a b = CONST
Equal Error Variances: PSA0 - PSA12
Path Diagram
End of Problem
```

There are two latent variables a and b in the model. They represent the intercept and slope of the patients' linear growth curves. The objective is to estimate the mean vector and covariance matrix of a and b and the error variance of the PSA measures. The error variance is assumed to be the same at all occasions.

In the current example, a and b are latent variables, and the line in the input file **psavar1a.spl**

a b = CONST

specifies that the means of a and b should be estimated.

The output gives the following information

```
--------------------------------
               EM Algorithm for missing Data:
              --------------------------------

           Number of different missing-value patterns=       14
           Effective sample size:      100

           Convergence of EM-algorithm in     9 iterations
           -2 Ln(L) =     1997.49237
           Percentage missing values=  13.40
```

The EM algorithm is first used to estimate a saturated model where both the mean vector and covariance matrix are unconstrained. This also gives the value $-2\ln(L) = 1997.492$.

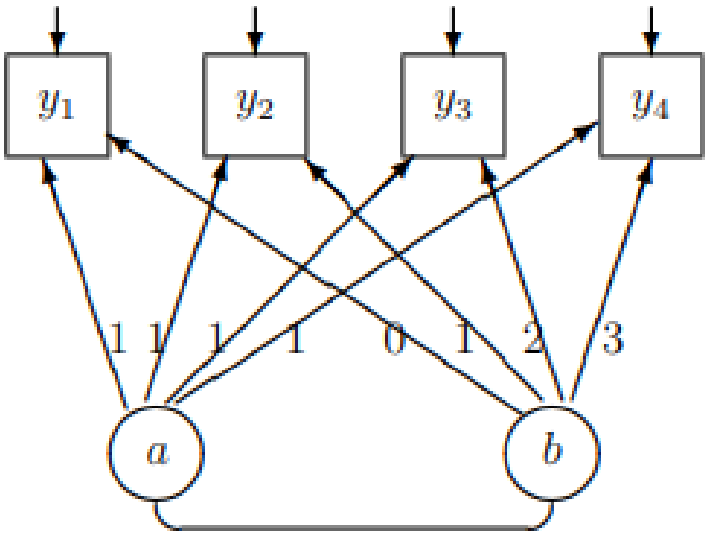


Figure 2: Path Diagram for a Linear Curve Model with Four Occasions

These are used to obtain starting values for the FIML method. After convergence the FIML method gives the following information about the fit of the model.

```
  Global Goodness of Fit Statistics, FIML case


               -2ln(L) for the saturated model =        1997.492
               -2ln(L) for the fitted model    =        2008.601

 Degrees of Freedom = 14
 Full Information ML Chi-Square                          11.108 (P = 0.6775)
 Root Mean Square Error of Approximation (RMSEA)         0.0
 90 Percent Confidence Interval for RMSEA                (0.0 ; 0.0775)
 P-Value for Test of Close Fit (RMSEA < 0.05)            0.844
```

The FIML estimates of the model parameters are given as

```
Covariance Matrix of Independent Variables

                a          b
          --------   --------
       a     30.899
            (4.612)
              6.700

       b      0.302      0.004
            (0.108)    (0.005)
              2.811      0.728


        Mean Vector of Independent Variables

                a          b
```

```
--------    --------
  31.934      -0.742
 (0.571)     (0.019)
  55.927     -39.871
```

The conclusions from this analysis are

- The average initial PSA value is 31.9 with a variance of 30.9.
- Thus, the initial PSA value varies considerably from patient to patient
- The effect of treatment is highly significant.
- The PSA value decreases by 0.7 per quarter (0.23 per year) and this rate of decrease is the same for all patients.

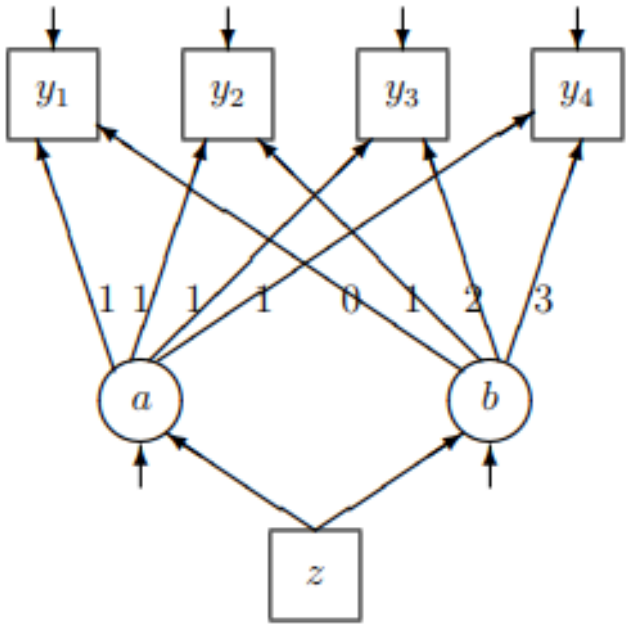


Figure 3: The Linear Curve Model with Covariate

To estimate the model in Figure 3 one can just add Age on the lines for *a* and *b*. The SIMPLIS syntax file is **psavar2a.spl**:

```
Linear Growth Curve with Covariate for psavar Data
Raw Data from File psavar.LSF
Latent Variables: a b
Relationships
PSA0 = 1*a 0*b
PSA3 = 1*a 3*b
PSA6 = 1*a 6*b
PSA9 = 1*a 9*b
PSA12 = 1*a 12*b
a b = CONST Age
Let the Errors on a and b correlate
Equal Error Variances: PSA0 - PSA12
Path Diagram
End of Problem
```

However, since we already know that all patients have the same slope *b*, it is not meaningful to predict *b* from Age. Thus instead of the line

```
a b = CONST Age
```

one should use (see file **psavar2aa.spl**)

```
a = CONST Age b = CONST
0*Age
```

The prediction equation for the intercept *a* is estimated as

```
        a = 15.288 + 0.300*Age, Errorvar.= 25.817, R² = 0.177
 Standerr  (3.691)  (0.0659)              (3.891)
 Z-values   4.141    4.555                 6.634
 P-values   0.000    0.000                 0.000
```

Thus, the intercept *a* depends on age. The intercept increases by 0.30 per year of age, on average.

There is an alternative method of estimation, based on the same two assumptions. One can use multiple imputation to obtain a complete data set and then analyze this by maximum likelihood or robust maximum likelihood method. Since the sample size *N* = 100 is small it is best to use maximum likelihood.

For the model in the last analysis the SIMPLIS syntax will be (see file **psavar3a.spl**):

```
!Linear Model with Covariate for psavar Data
!Estimated by ML using Multiple Imputation
Raw Data from File psavar.lsf
Multiple Imputation with MC
Latent Variables: a b
Relationships
PSA0  = 1*a 0*b
PSA3  = 1*a 3*b
PSA6  = 1*a 6*b
PSA9  = 1*a 9*b
PSA12 = 1*a 12*b
a = CONST Age
b = CONST 0*Age
Let the Errors of a and b correlate
Equal Error Variances: PSA0 - PSA12
Path Diagram
End of Problem
```

The only difference between this input file and **psavar2aa.spl** is the line

```
Multiple Imputation
```

which has been added. The output gives the following estimated equation for *a*:

```
        a = 15.000 + 0.306*Age, Errorvar.= 26.020, R² = 0.183
```

```
 Standerr  (3.570)  (0.0637)              (3.849)
 Z-values   4.202    4.798                 6.761
 P-values   0.000    0.000                 0.000
```

which is very similar to previous results.

An advantage of this approach is the one can get more measures of goodness of fit:

```
  Log-likelihood Values

                        Estimated Model          Saturated Model
                        ---------------          ---------------
 Number of free parameters(t)         9                       27
 -2ln(L)                        1787.430                1764.336
 AIC (Akaike, 1974)*            1805.430                1818.336
 BIC (Schwarz, 1978)*           1828.877                1888.675

*LISREL uses AIC= 2t - 2ln(L) and BIC = tln(N)- 2ln(L)


                           Goodness-of-Fit Statistics

 Degrees of Freedom for (C1)-(C2)                      18
 Maximum Likelihood Ratio Chi-Square (C1)              23.095 (P = 0.1870)
 Due to Covariance Structure                           0.0
 Due to Mean Structure                                 0.0
 Browne's (1984) ADF Chi-Square (C2_NT)                0.0 (P = 1.0000)

 Estimated Non-centrality Parameter (NCP)              5.095
 90 Percent Confidence Interval for NCP                (0.0 ; 21.634)

 Minimum Fit Function Value                            0.231
 Population Discrepancy Function Value (F0)            0.0509
 90 Percent Confidence Interval for F0                 (0.0 ; 0.216)
 Root Mean Square Error of Approximation (RMSEA)       0.0532
 90 Percent Confidence Interval for RMSEA              (0.0 ; 0.110)
 P-Value for Test of Close Fit (RMSEA < 0.05)          0.426
```