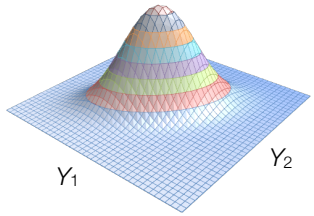
Maximum Likelihood Estimation with Incomplete Data

Multivariate Normal Density Function

L_i gives the relative probability that the set of scores in **Y** came from a multivariate normal distribution with a particular mean vector and covariance matrix



$$L_i = \frac{1}{\sqrt{2\pi}^{k/2} \mid \mathbf{\Sigma} \mid^{.5}} exp\left(-\frac{(\mathbf{Y}_i - \boldsymbol{\mu})' \mathbf{\Sigma}^{-1} (\mathbf{Y}_i - \boldsymbol{\mu})}{2}\right)$$

Log Likelihood Comparison

Complete data log likelihood

$$lnL = \sum_N ln\left[\frac{1}{\sqrt{2\pi}^{k/2} \mid \mathbf{\Sigma} \mid^{.5}} exp\left(-\frac{(\mathbf{Y}_i - \boldsymbol{\mu})' \mathbf{\Sigma}^{-1} (\mathbf{Y}_i - \boldsymbol{\mu})}{2}\right)\right]$$

Missing data log likelihood

$$lnL = \sum_N ln\left[\frac{1}{\sqrt{2\pi}^{k/2} \mid \mathbf{\Sigma}_i \mid^{.5}} exp\left(-\frac{(\mathbf{Y}_i - \boldsymbol{\mu}_i)' \mathbf{\Sigma}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i)}{2}\right)\right]$$

Missing-Data Log Likelihood

The missing data log likelihood has an *i* subscript on the mean vector and covariance matrix

The subscript indicates that the elements in these matrices vary across missing data patterns

The fit of a case's data to the parameters is evaluated using only those parameters for which there is data

Missing-Data Log Likelihood, Continued

$$lnL = \sum_{N} ln \left[\frac{1}{\sqrt{2\pi}^{k/2} \mid \boldsymbol{\Sigma}_i \mid^{.5}} exp \left(- \frac{(\mathbf{Y}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i)}{2} \right) \right]$$

Mean vector elements corresponding to observed scores

Observed scores for case *i*

Covariance matrix elements corresponding to observed scores

Motivating Example

Number of years smoking (*X*) and number of cigarettes smoked (*Y*)

Estimate means and variance-covariance matrix

All cases contribute information if they have at least one observation

Years (X)	Cigs (Y)
7	9
8	NA
1	11
4	3
6	10
8	5
8	7
10	NA
15	NA
5	11
9	12
11	11
14	NA
13	19
12	NA
11	8
10	13
10	8
7	10
11	10

Mahalanobis Distance Computations

Complete cases

$$z_i^2 = (\mathbf{Y}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i)$$

$$= \begin{pmatrix} X_i \\ Y_i \end{pmatrix} - \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}' \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{YX} & \sigma_Y^2 \end{pmatrix}^{-1} \begin{pmatrix} X_i \\ Y_i \end{pmatrix} - \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$$

Incomplete cases

$$z_i^2 = (X_i - \mu_X)' (\sigma_X^2)^{-1} (X_i - \mu_X) = \frac{(X_i - \mu_X)^2}{\sigma_X^2}$$

Sample Log Likelihood

$$lnL = \sum_{i=1}^{n_{(com)}} ln \left[\frac{1}{\sqrt{2\pi} \mid \boldsymbol{\Sigma} \mid^{.5}} exp \left(- \frac{(\mathbf{Y}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_i - \boldsymbol{\mu})}{2} \right) \right]$$

$$+ \sum_{j=1}^{n_{(mis)}} ln \left[\frac{1}{\sqrt{2\pi\sigma_X^2}} exp \left(- \frac{(X_i - \mu_X)^2}{2\sigma_X^2} \right) \right]$$

Log likelihood for $n_{(com)}$ complete cases

Log likelihood for $n_{(mis)}$ incomplete cases

Estimation Steps

Step 0: Generate starting values for all parameters

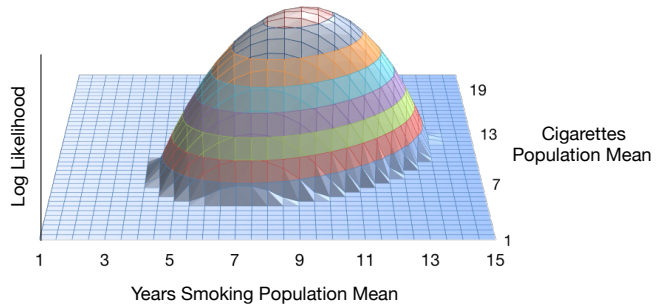
Step 1: Generate temporary “imputations” based on the current model parameters

Step 2: Update parameters based on the observed data and temporary imputations

Repeat Steps 1 and 2 until the estimates from Step 2 no longer change from one iteration to the next

Log Likelihood Surface

The log likelihood surface reflects the probability of the data for different combinations of the population means



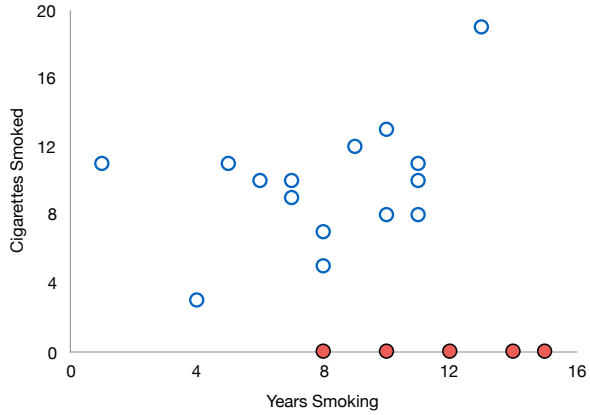
How Do the Incomplete Cases Help Estimation?

Maximum likelihood uses all available data to estimate the parameters

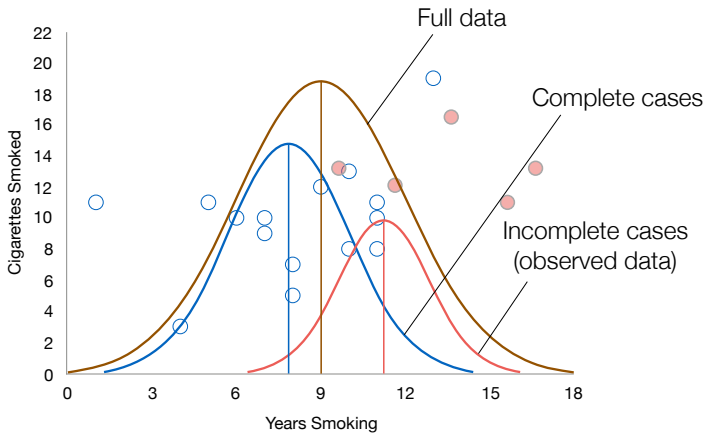
The procedure can be viewed as implicit imputation because the observed data imply plausible values for the missing scores

Multivariate normality is key because it implies a distribution of plausible scores for the missing data

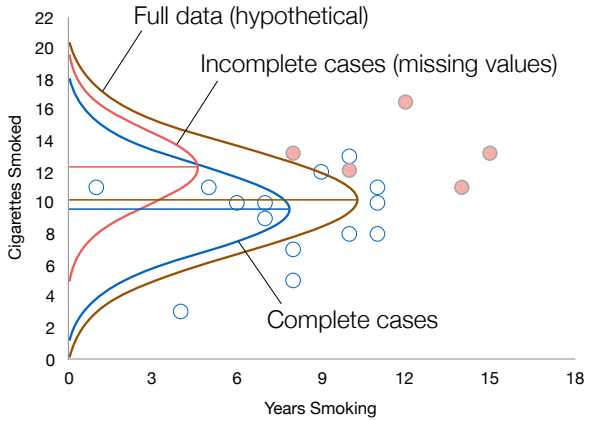
Observed Data



Distribution of Years Smoking



Distribution of Cigarettes Smoked



Source of Deletion Bias

The MAR mechanism functions such that more years smoking increases the likelihood of nonresponse

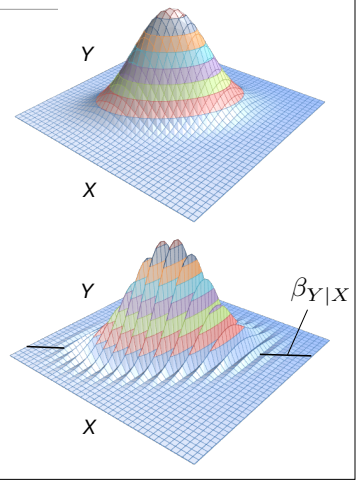
Excluding incomplete cases removes observations from the upper tail of both distributions because the variables are positively correlated

The resulting means and variances are too low

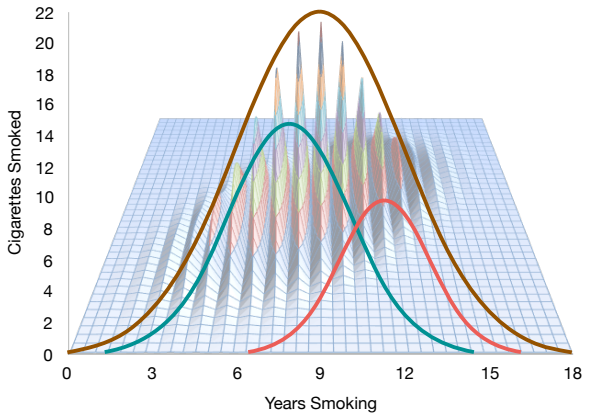
Bivariate Normal Distribution

The bivariate normal distribution can be viewed as a collection of conditionals

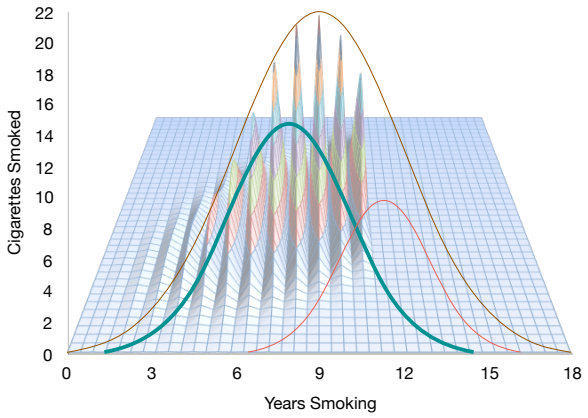
Each slice represents the distribution of *Y* (cigarettes smoked) for a particular value of *X* (years smoking)



Conditionals Based on Full Observed Data



Conditionals Based on Complete Cases

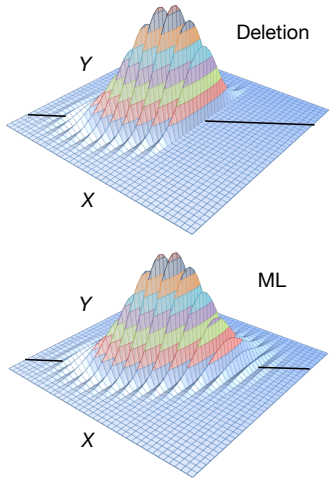


Estimation Adjustment

Including cases with complete data on only Y_1 (years smoking) increases the Y_1 variance

Adding variance implies high values for the missing Y_2 scores

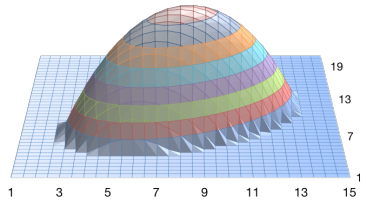
The estimator increases the mean and variance of Y_2



Comparison of Log Likelihood Surfaces

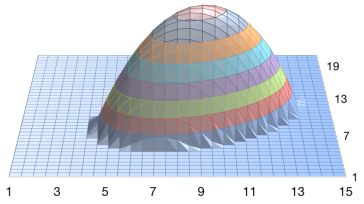
The ML log likelihood is steeper (small standard errors) and its maximum is centered at different values

Complete cases (N = 15)



Years Smoking
Population Mean

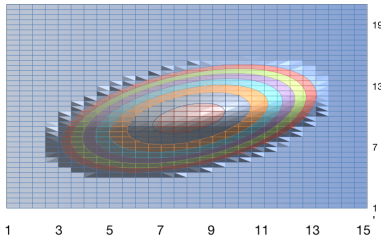
Maximum likelihood (N = 20)



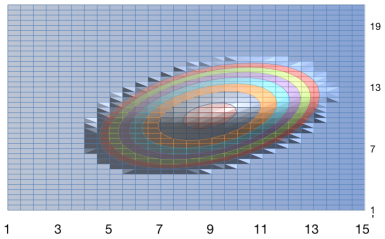
Years Smoking
Population Mean

Top View

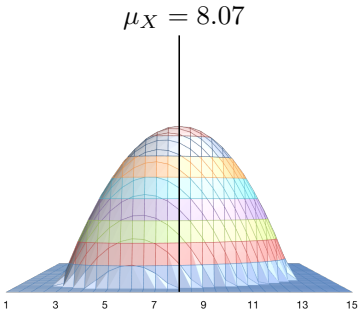
Complete cases (*N* = 15)



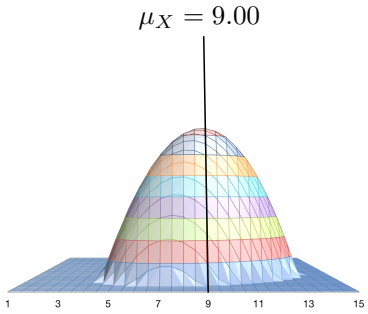
Maximum likelihood (*N* = 20)



Years Smoking Mean Estimates

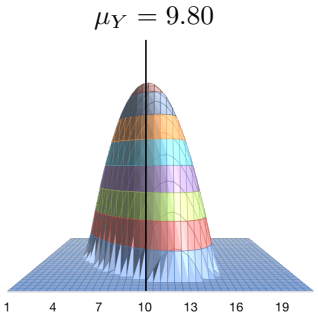


Complete cases (*N* = 15)

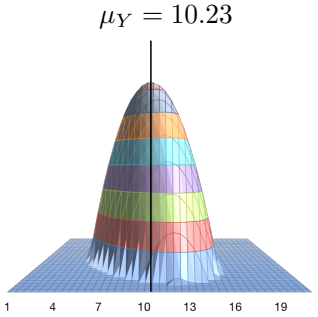


Maximum likelihood (*N* = 20)

Years Smoking Mean Estimates



Complete cases (*N* = 15)



Maximum likelihood (*N* = 20)

Ex10.1.inp
Mplus Analysis Script

```
DATA:
file = smoking.csv;
VARIABLE:
names = id quitmeth male age years cigs heavycig
 efficacy stress;
usevariables = years cigs;
missing = all(-99);
MODEL:
years cigs;
years with cigs;
OUTPUT:
standardized(stdyx);
```

Mplus Analysis Output

```
SUMMARY OF ANALYSIS

Number of groups                                 1
Number of observations                          20

Number of dependent variables                    2
Number of independent variables                  0
Number of continuous latent variables            0
```

Mplus Analysis Output

```
MODEL RESULTS

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

YEARS    WITH
    CIGS               5.228      3.506      1.491      0.136

Means
    YEARS              9.000      0.752     11.973      0.000
    CIGS              10.232      0.943     10.849      0.000

Variances
    YEARS             11.301      3.574      3.162      0.002
    CIGS              12.967      4.858      2.669      0.008
```

Mplus Analysis Output

```
STANDARDIZED MODEL RESULTS

STDYX Standardization

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

YEARS    WITH
    CIGS               0.432      0.224      1.924      0.054

Means
    YEARS              2.677      0.479      5.592      0.000
    CIGS               2.841      0.554      5.124      0.000

Variances
    YEARS              1.000      0.000    999.000    999.000
    CIGS               1.000      0.000    999.000    999.000
```

Analysis Example

Motivating Example

Analysis involving number of years smoking (Years), number of cigarettes smoked per week (Cigs), and self-efficacy to quit smoking (Efficacy)

Imputation assumes that values are missing at random

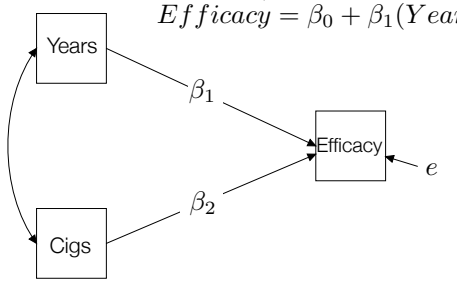
What does MAR require in this context?

Years	Cigs	Efficacy
7	9	NA
8	NA	NA
1	11	16
4	3	21
6	10	17
8	5	10
8	7	13
10	NA	10
15	NA	11
5	11	13
9	12	11
11	11	16
14	NA	10
13	19	9
12	NA	5
11	8	7
10	13	10
10	8	NA
7	10	7
11	10	6

Analysis Model

The analysis model is a multiple regression predicting self-efficacy to quit based on years smoking and number of cigarettes smoked

$$Efficacy = \beta_0 + \beta_1(Years) + \beta_2\,(Cigs) + e$$



Mplus Regression Script

```
DATA:
file = smoking.csv;
VARIABLE:
names = id quitmeth male age years
  cigs heavycig efficacy stress;
usevariables = efficacy years cigs;
missing = all(-99);
MODEL:
efficacy on years (b1)
  cigs (b2);
MODEL TEST:
b1 = 0; b2 = 0;
OUTPUT:
standardized(stdyx);
```

Mplus Analysis Output

```
*** 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:  5
*** WARNING
  Data set contains cases with missing on all variables except
  x-variables.  These cases were not included in the analysis.
  Number of cases with missing on all variables except x-variables:  2
   2 WARNING(S) FOUND IN THE INPUT INSTRUCTIONS
```

Mplus Analysis Output

```
SUMMARY OF ANALYSIS

Number of groups                                 1
Number of observations                          13

Number of dependent variables                    1
Number of independent variables                  2
Number of continuous latent variables            0
```

Regression Log Likelihood

Regression models do not impose distributional assumptions on predictor variables (predictors are "fixed")

$$lnL = \sum_{N} ln \left[\frac{1}{\sqrt{2\pi\sigma_e^2}} exp \left(-.5 \frac{(Y_i - \mathbf{X}\boldsymbol{\beta})^2}{\sigma_e^2} \right) \right]$$

Predictors are attached to the parameters in the log likelihood function and effectively function as constants

Latent Variable Framework

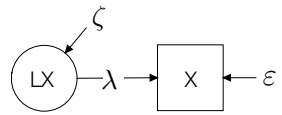
The latent variable framework provides a natural mechanism for handling incomplete predictors

Some software packages (e.g., LISREL, EQS, CALIS) automatically treat all manifest variables as normally distributed, while others can but do not by default (e.g., Mplus, Laavan)

Single-Indicator Measurement Model

A regression weight (factor loading) links the latent variable (LX) to the manifest variable

The latent variable has a mean and variance, and the manifest variable has an intercept and residual variance

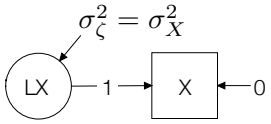


Parameter Constraints

Parameter constraints produce a latent variable with the same mean and variance as its indicator

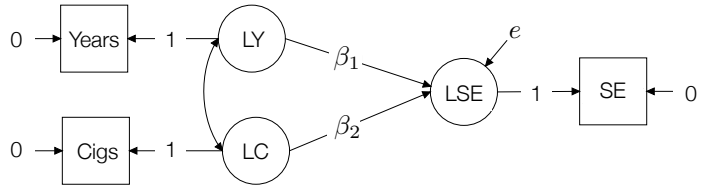
Loading = 1, residual variance = 0, intercept = 0

The manifest variable is now an outcome that appears in the **Y** vector of the log likelihood



Latent Variable Regression Model

All manifest variables are treated as normally distributed outcomes, thus allowing for missing data handling



Ex10.2.inp
Mplus Analysis Script

```
DATA:
file = smoking.csv;
VARIABLE:
names = id quitmeth male age years
  cigs heavycig efficacy stress;
usevariables = efficacy years cigs;
missing = all(-99);
MODEL:
years cigs; ! List all predictors, even if complete;
efficacy on years (b1)
  cigs (b2);
MODEL TEST:
b1 = 0; b2 = 0;
OUTPUT:
standardized(stdyx);
```

Mplus Analysis Output

```
SUMMARY OF ANALYSIS

Number of groups                                 1
Number of observations                          20

Number of dependent variables                    2
Number of independent variables                  0
Number of continuous latent variables            0
```

Mplus Analysis Output

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
EFFICACY ON				
YEARS	-0.637	0.254	-2.505	0.012
CIGS	-0.106	0.267	-0.397	0.691
CIGS WITH				
YEARS	5.073	3.498	1.450	0.147
Means				
YEARS	9.000	0.752	11.973	0.000
CIGS	10.214	0.939	10.878	0.000

Mplus Analysis Output

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
...				
Intercepts				
EFFICACY	18.207	2.783	6.541	0.000
Variances				
YEARS	11.300	3.573	3.162	0.002
CIGS	12.823	4.770	2.688	0.007
Residual Variances				
EFFICACY	11.069	3.807	2.907	0.004

Mplus Analysis Output

STANDARDIZED MODEL RESULTS

STDYX Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
EFFICACY ON				
YEARS	-0.527	0.188	-2.811	0.005
CIGS	-0.093	0.234	-0.400	0.689
CIGS WITH				
YEARS	0.421	0.229	1.838	0.066
Means				
YEARS	2.677	0.479	5.592	0.000
CIGS	2.852	0.554	5.146	0.000

Mplus Analysis Output

STANDARDIZED MODEL RESULTS

STDYX Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Intercepts				
EFFICACY	4.485	0.702	6.385	0.000
Variances				
YEARS	1.000	0.000	999.000	999.000
CIGS	1.000	0.000	999.000	999.000
Residual Variances				
EFFICACY	0.672	0.177	3.798	0.000