4. Maximum Likelihood Missing Data Handling

Simple Regression Analysis

Simple regression where Y has missing values

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = E(Y|X) + \varepsilon_i$$
$$Y_i \sim N\left(E(Y|X), \sigma_\varepsilon^2\right)$$

$E(Y|X)$ is a predicted value from the regression

Leader-Member Exchange Data

Work-related data for 630 employees nested in 105 different workgroups

The data include work-related variables such as employee empowerment, job satisfaction, turnover intentions, employee-supervisor relationship quality, organizational climate

lmxquality.dat

Variable	Name	Missing %	Scaling
Employee identifier	EMPLOYEE	0	Integer index
Team identifier	TEAM	0	Integer index
Turnover intentions	TURNOVER	5.1	0 = intend to stay, 1 = intend to leave
Gender	MALE	0	0 = female, 1 = male
Employee empowerment	EMPOWER	16.2	Continuous
Leader-member exchange	LMXQUALITY	4.1	Continuous
Job satisfaction	JOBSAT	4.8	7-point ordinal scale
Organizational (team) climate	CLIMATE	9.5	Continuous
Organization size	ORGSIZE	5.7	6-point ordinal scale

Substantive Example

Employee empowerment regressed on employee-supervisor relationship quality

$$EMPOWER_i = \beta_0 + \beta_1(LMX_i) + \varepsilon_i$$

Empowerment score are normally distributed around the regression line

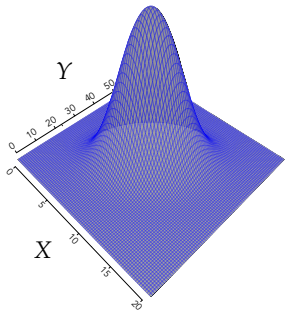
Multivariate Normality

All incomplete variables need a distribution

Maximum likelihood software generally assumes that variables are multivariate normal

Mixtures of categorical and continuous variables are best handled with Bayesian estimation or multiple imputation

Bivariate Normal Distribution



The likelihood is the height of the bivariate normal distribution at the intersection of two scores

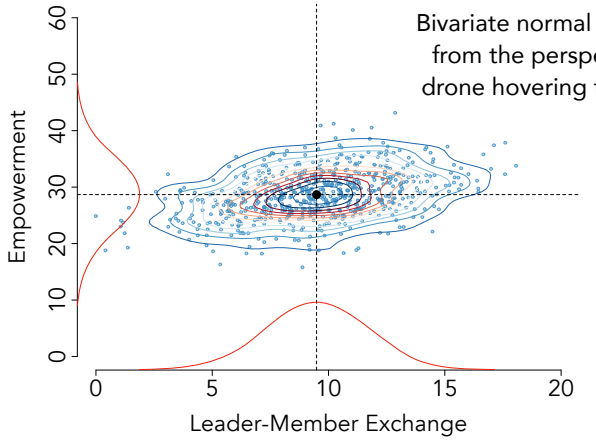
$$f(\mathbf{y}_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{v}{\sqrt{2\pi|\boldsymbol{\Sigma}|}} \times \exp\left\{-\frac{1}{2}\underbrace{(\mathbf{y}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu})}_{\text{Squared } z\text{-score}}\right\}$$

Closer Look At The Squared z-Score

The likelihood (relative probability) increases as the standardized distance between X and Y and the center of the distribution decreases

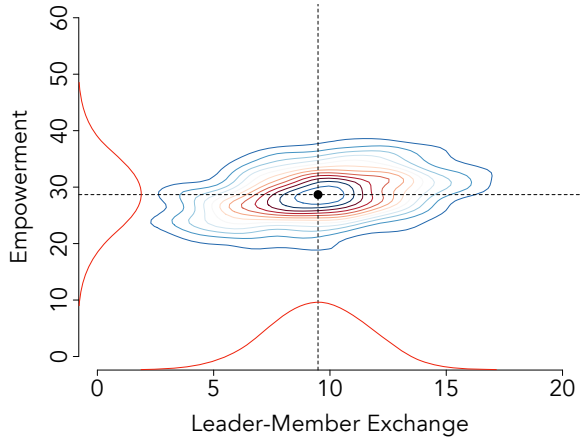
$$f(\mathbf{y}_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \text{scaling term} \times \exp\left\{-\frac{1}{2}\begin{pmatrix}Y_i\\X_i\end{pmatrix}-\begin{pmatrix}\mu_Y\\\mu_X\end{pmatrix}^T\begin{pmatrix}\sigma^2_Y & \sigma_{YX}\\\sigma_{XY} & \sigma^2_X\end{pmatrix}^{-1}\begin{pmatrix}Y_i\\X_i\end{pmatrix}-\begin{pmatrix}\mu_Y\\\mu_X\end{pmatrix}\right\}$$

Scatterplot With Contours



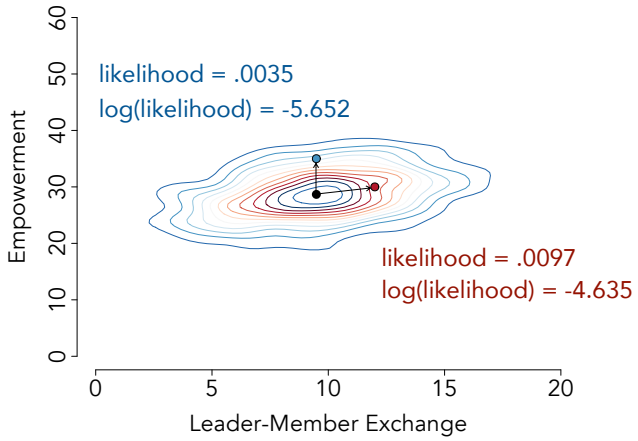
9

Contour Plot



10

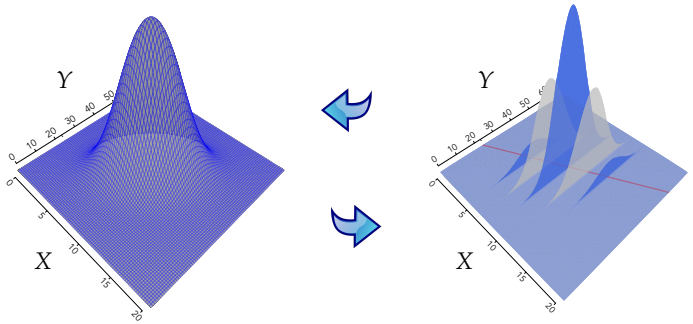
Likelihoods For Two Observations



11

Bivariate Normal As Regression

A bivariate normal distribution can be expressed as a regression, and vice versa



12

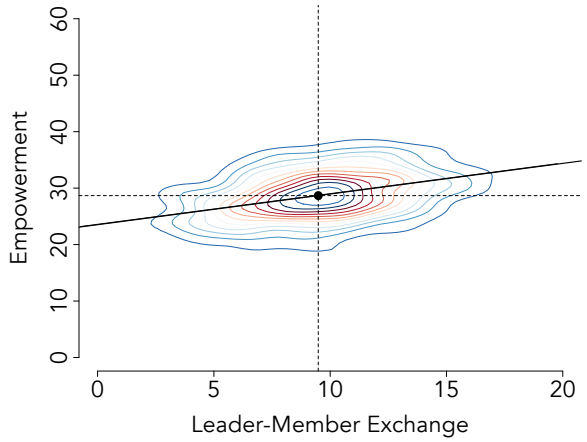
Squared z-Score As Regression

The squared z-score still captures the standardized distance between a pair of scores and the center of the distribution

$$f(\mathbf{y}_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \ldots \quad \begin{matrix}\mu_Y \\ \updownarrow \\ \mu_X\end{matrix} \quad \begin{matrix}\sigma^2_Y \\ \updownarrow \\ \sigma_{YX}\end{matrix}$$

$$\times \exp\left\{-\frac{1}{2}\left(\begin{pmatrix}Y_i\\X_i\end{pmatrix}-\begin{pmatrix}\beta_0+\beta_1\mu_X\\ \mu_X\end{pmatrix}\right)^T \begin{pmatrix}\beta_1^2\sigma_X^2+\sigma_\varepsilon^2 & \beta_1\sigma_X^2\\ \beta_1\sigma_X^2 & \sigma_X^2\end{pmatrix}^{-1}\left(\begin{pmatrix}Y_i\\X_i\end{pmatrix}-\begin{pmatrix}\beta_0+\beta_1\mu_X\\ \mu_X\end{pmatrix}\right)\right\}$$

Contour Plot With Regression Line



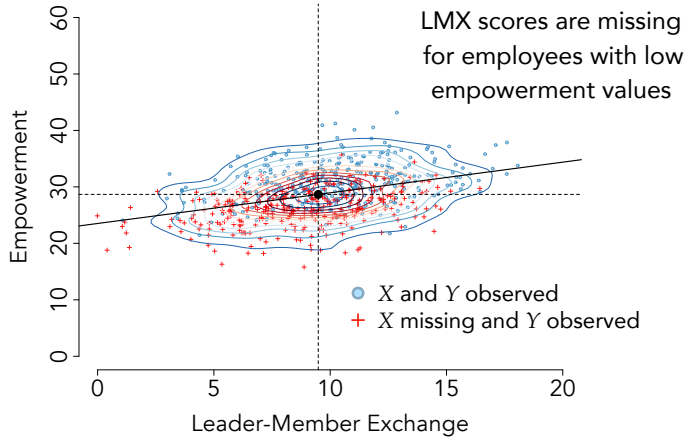
Missing Data Handling

Maximum likelihood uses all available data

Each participant contributes their observed scores, data records need not be complete

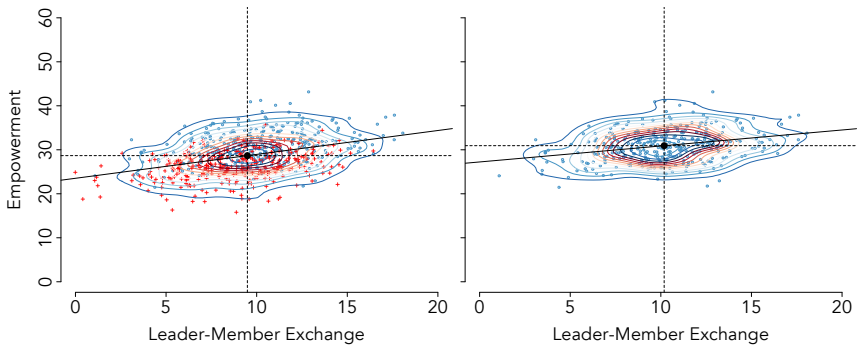
ML is sometimes called "implicit imputation" because the observed data and the normal distribution imply values for the missing scores

Example: MAR Mechanism



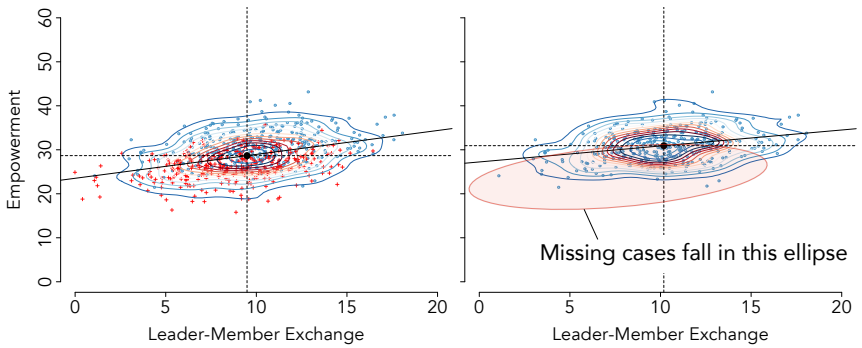
Incomplete vs. Complete Data

The complete cases are systematically different, with a flatter regression line and less variation



Incomplete vs. Complete Data

The complete cases are systematically different, with a flatter regression line and less variation



Individual Likelihood Revisited

Participants with complete data (the ● symbols) contribute two scores, and their squared z-values are computed from two deviation scores

$$f(\mathbf{y}_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \ldots$$

$$\times \exp\left\{-\frac{1}{2}\left(\begin{pmatrix}Y_i\\X_i\end{pmatrix}-\begin{pmatrix}\beta_0+\beta_1\mu_X\\\mu_X\end{pmatrix}\right)^T\begin{pmatrix}\beta_1^2\sigma_X^2+\sigma_\varepsilon^2 & \beta_1\sigma_X^2\\\beta_1\sigma_X^2 & \sigma_X^2\end{pmatrix}^{-1}\left(\begin{pmatrix}Y_i\\X_i\end{pmatrix}-\begin{pmatrix}\beta_0+\beta_1\mu_X\\\mu_X\end{pmatrix}\right)\right\}$$

Maximum Likelihood Missing Data

Maximum likelihood uses all available data, including the Y scores for cases missing X

Squared z-values are computed using only the parameters for which there is data

The partial data records steer the estimation routine to a more accurate set of estimates

Log Likelihood Contribution With Missing X

Eliminate terms corresponding to missing variable

$$f(\mathbf{y}_i|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \text{scaling term} \times \exp\left\{-\frac{1}{2}\left(\begin{pmatrix} Y_i \\ \cancel{X_i} \end{pmatrix} - \begin{pmatrix} \beta_0 + \beta_1\mu_X \\ \cancel{\mu_X} \end{pmatrix}\right)^T \begin{pmatrix} \beta_1^2\sigma_X^2 + \sigma_\varepsilon^2 & \cancel{\beta_1\sigma_X^2} \\ \cancel{\beta_1\sigma_X^2} & \cancel{\sigma_X^2} \end{pmatrix}^{-1} \left(\begin{pmatrix} Y_i \\ \cancel{X_i} \end{pmatrix} - \begin{pmatrix} \beta_0 + \beta_1\mu_X \\ \cancel{\mu_X} \end{pmatrix}\right)\right\}$$

Squared z-score is computed using only Y

$$f(\mathbf{y}_i|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \text{scaling term} \times \exp\left\{-\frac{1}{2}\frac{\left(Y_i - (\beta_0 + \beta_1\mu_X)\right)^2}{\beta_1^2\sigma_X^2 + \sigma_\varepsilon^2}\right\}$$

Log Likelihood Contribution With Missing Y

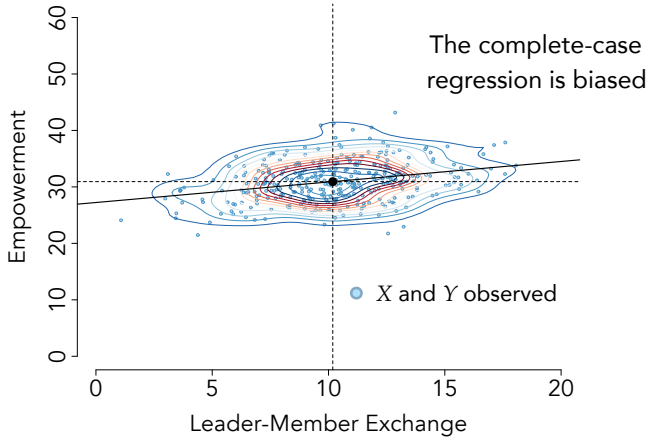
Eliminate terms corresponding to missing variable

$$f(\mathbf{y}_i|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \text{scaling term} \times \exp\left\{-\frac{1}{2}\left(\begin{pmatrix} \cancel{Y_i} \\ X_i \end{pmatrix} - \begin{pmatrix} \cancel{\beta_0 + \beta_1\mu_X} \\ \mu_X \end{pmatrix}\right)^T \begin{pmatrix} \beta_1^2\sigma_X^2 + \sigma_\varepsilon^2 & \cancel{\beta_1\sigma_X^2} \\ \cancel{\beta_1\sigma_X^2} & \sigma_X^2 \end{pmatrix}^{-1} \left(\begin{pmatrix} \cancel{Y_i} \\ X_i \end{pmatrix} - \begin{pmatrix} \cancel{\beta_0 + \beta_1\mu_X} \\ \mu_X \end{pmatrix}\right)\right\}$$

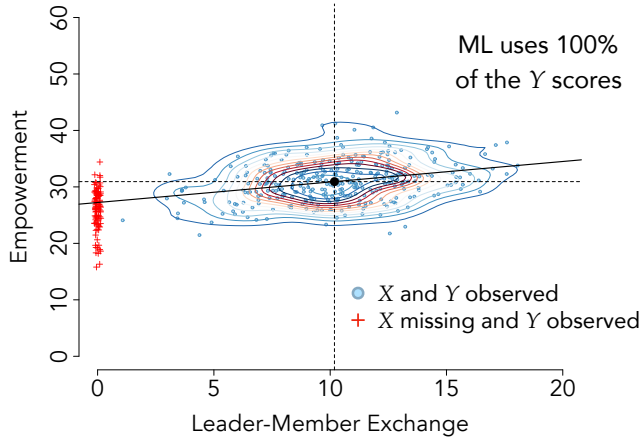
Squared z-score is computed using only X

$$f(\mathbf{y}_i|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \text{scaling term} \times \exp\left\{-\frac{1}{2}\frac{(X_i - \mu_X)^2}{\sigma_X^2}\right\}$$

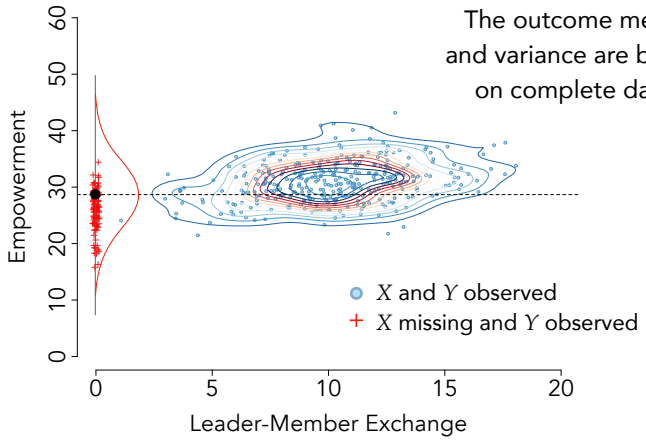
Complete-Case Analysis



Leveraging The Partial Data Records

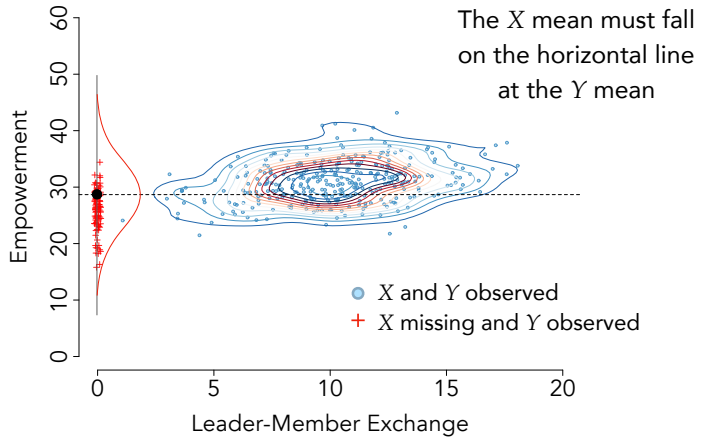


How Do Partial Data Records Help?



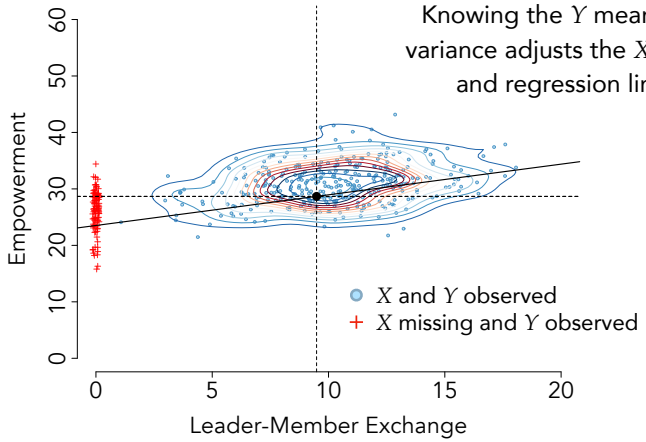
25

How Do Partial Data Records Help?



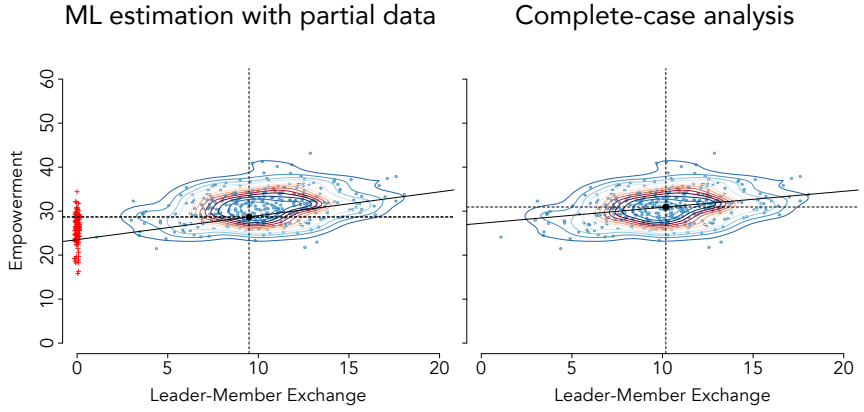
26

How Do Partial Data Records Help?



27

Visual Comparison Of Estimates



28

Maximum Likelihood Estimates

Results based on all available data from 250 participants

Parameter	Est.	*SE*	z	p
Intercept	22.65	0.59	38.21	< .001
LMX slope	0.62	0.06	9.93	< .001

Interpretations

Same as ordinary least squares!

The intercept is the expected (predicted) value when leader-member exchange (employee-supervisor relationship quality) equals zero

A one-unit increase in leader-member exchange (relationship quality) increases employee empowerment by .62, on average ($p < .001$)

Mplus Maximum Likelihood Descriptives Script

```
DATA:
file = lmxquality.dat;
VARIABLE:
names = employee team turnover male empower lmxquality
  jobsat climate orgsize;
missing = all(999);
usevariables = empower lmxquality;
MODEL:
empower  with  lmxquality;
OUTPUT:
stdyx;
```


Mplus Output

```
SUMMARY OF ANALYSIS

Number of groups                                         1
Number of observations                                 630

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

...

MODEL FIT INFORMATION

Number of Free Parameters                        5

Loglikelihood

          H0 Value                       -3007.913
          H1 Value                       -3007.913
```

Mplus Output

```
MODEL RESULTS

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

EMPOWER  WITH
   LMXQUALITY        5.622      0.649      8.667      0.000

Means
   EMPOWER          28.585      0.192    149.210      0.000
   LMXQUALITY        9.631      0.122     78.650      0.000

Variances
   EMPOWER          19.432      1.214     16.001      0.000
   LMXQUALITY        9.124      0.524     17.421      0.000
```


Mplus Output

```
STANDARDIZED MODEL RESULTS

STDYX Standardization

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

EMPOWER  WITH
   LMXQUALITY        0.422      0.038     11.120      0.000

Means
   EMPOWER           6.485      0.203     31.884      0.000
   LMXQUALITY        3.189      0.100     31.873      0.000

Variances
   EMPOWER           1.000      0.000    999.000    999.000
   LMXQUALITY        1.000      0.000    999.000    999.000
```


Mplus Maximum Likelihood Script

DATA:
file = lmxquality.dat;
VARIABLE:
names = employee team turnover male empower lmxquality
  jobsat climate orgsize;
missing = all(999);
usevariables = empower lmxquality;
MODEL:
empower on lmxquality;
OUTPUT:
stdyx;

No distribution for the incomplete predictor!

Mplus 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:  26
*** 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:  102
   3 WARNING(S) FOUND IN THE INPUT INSTRUCTIONS

SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                         502

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

Mplus Maximum Likelihood Script

```
DATA:
file = lmxquality.dat;
VARIABLE:
names = employee team turnover male empower lmxquality
  jobsat climate orgsize;
missing = all(999);
usevariables = empower lmxquality;
MODEL:
lmxquality;  <-------- Normal distribution for the incomplete predictor
empower  on  lmxquality;
OUTPUT:
stdyx;
```


Mplus Output

```
SUMMARY OF ANALYSIS

Number of groups                                 1
Number of observations                         630

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

...

MODEL FIT INFORMATION

Number of Free Parameters                        5

Loglikelihood

          H0 Value                       -3007.913
          H1 Value                       -3007.913
```


Mplus Output

```
MODEL RESULTS

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

 EMPOWER  ON
    LMXQUALITY       0.616      0.062      9.934      0.000

 Means
    LMXQUALITY       9.631      0.122     78.649      0.000

 Intercepts
    EMPOWER         22.650      0.593     38.209      0.000

 Variances
    LMXQUALITY       9.124      0.524     17.421      0.000

 Residual Variances
    EMPOWER         15.967      0.989     16.149      0.000
```


Mplus Output

```
STANDARDIZED MODEL RESULTS

STDYX Standardization

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

 EMPOWER  ON
    LMXQUALITY       0.422      0.038     11.121      0.000

...

R-SQUARE

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

    EMPOWER            0.178      0.032      5.560      0.000
```

R Mplus Maximum Likelihood Descriptives Script

```
# load lavaan sem package
library(lavaan)

# read data and assign variable names
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
lmxdata <- read.table(paste0(getwd(), "/lmxquality.dat"))
names(lmxdata) <- c("employee", "team", "turnover", "male", "empower",
  "lmxquality", "jobsat", "climate", "orgsize")

# missing data code
lmxdata[lmxdata == 999] <- NA
```


R Output

```
  Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                          5

  Number of observations                           630
  Number of missing patterns                         3

...

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -3007.913
  Loglikelihood unrestricted model (H1)      -3007.913
```


R Output

```
Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  empower ~~
    lmxquality        5.622    0.649    8.667    0.000    5.622    0.422

Intercepts:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    empower          28.585    0.192  149.210    0.000   28.585    6.485
    lmxquality        9.631    0.122   78.650    0.000    9.631    3.188

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    empower          19.432    1.214   16.001    0.000   19.432    1.000
    lmxquality        9.124    0.524   17.421    0.000    9.124    1.000
```


R Mplus Maximum Likelihood Script

```
# specify model
model <- '
empower ~~ lmxquality
'

# fit model and summarize estimates
analysis <- sem(model, lmxdata, missing = 'fiml', fixed.x = F, meanstructure = T,
  mimic = "Mplus")
summary(analysis, fit.measures = TRUE, rsquare = T, standardize = T)
```

R Mplus Maximum Likelihood Script

```
# load lavaan sem package
library(lavaan)

# read data and assign variable names
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
lmxdata <- read.table(paste0(getwd(), "/lmxquality.dat"))
names(lmxdata) <- c("employee", "team", "turnover", "male", "empower",
  "lmxquality", "jobsat", "climate", "orgsize")

# missing data code
lmxdata[lmxdata == 999] <- NA
```


R Mplus Maximum Likelihood Script

```
# specify model
model <- '
empower ~ lmxquality
lmxquality ~~ lmxquality
'

# fit model and summarize estimates
analysis <- sem(model, lmxdata, missing = 'fiml', fixed.x = F, meanstructure = T,
  mimic = "Mplus")
summary(analysis, fit.measures = TRUE, rsquare = T, standardize = T)
```


R Output

```
  Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                          5

  Number of observations                           630
  Number of missing patterns                         3

...

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -3007.913
  Loglikelihood unrestricted model (H1)      -3007.913
```


R Output

```
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  empower ~
    lmxquality        0.616    0.062    9.934    0.000    0.616    0.422

Intercepts:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   .empower          22.650    0.593   38.209    0.000   22.650    5.138
    lmxquality        9.631    0.122   78.650    0.000    9.631    3.188

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    lmxquality        9.124    0.524   17.421    0.000    9.124    1.000
   .empower          15.967    0.989   16.149    0.000   15.967    0.822

R-Square:
                   Estimate
    empower           0.178
```

Multiple Regression Analysis

Multiple regression with two predictors

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i = E(Y|X) + \varepsilon_i$$
$$Y_i \sim N\left(E(Y|X), \sigma_\varepsilon^2\right)$$

Y is normally distributed around the regression line (predicted values) with constant variation

Chronic Pain Data

Pain-related data for 275 chronic pain patients

The data include psychological correlates of pain severity such as depression, pain interference with daily life, perceived control over pain, stress, and psychosocial disability

pain.dat

Variable	Name	Missing %	Scaling
Patient identifier	ID	0	Integer index
Gender	MALE	0	0 = female, 1 = male
Age	AGE	0	Continuous
Education level	EDUGROUP	0	1 = Some college or less, 2 = college, 3 = Post-BA
Work hours per week	WORKHRS	11.7	Continuous
Exercise	EXERCISE	1.7	8-point ordinal scale
Pain intensity rating	PAIN	7.3	1 = none/little, 2 = moderate, 3 = severe/very severe
Anxiety	ANXIETY	6.0	Continuous
Stress	STRESS	0	7-point ordinal scale
Perceived control over	CONTROL	0	Continuous
Pain interference with life	INTERFERE	13.3	Continuous
Depression	DEPRESS	13.3	Continuous
Psychosocial disability	DISABILITY	3.0	Continuous

Substantive Example

Gender (complete) and incomplete depression scores predict psychosocial disability

$$DISABILITY_i = \beta_0 + \beta_1(DEPRESS_i) + \beta_2(MALE_i) + \varepsilon_i$$

Disability measures pain's impact on emotional behaviors such as psychological autonomy and communication, emotional stability, etc.

Maximum Likelihood Estimates

Results based on all available data from 250 participants

Parameter	Est.	*SE*	z	p
Intercept	12.25	0.65	18.78	< .001
DEPRESS slope	0.32	0.04	7.87	< .001
MALE slope	-0.38	0.49	-0.79	0.433

Interpretations

Controlling for gender, a one-unit increase in depression predicts an increase in psychosocial disability of .32 ($p < .001$)

Controlling for depression, males have a psychosocial disability mean that is .38 points lower than that of females ($p = .43$)

Mplus Maximum Likelihood Script

```
DATA:
file = pain.dat;
VARIABLE:
names = id male age edugroup workhrs exercise pain anxiety
stress control
  interfere depress disability;
missing = all(999);
usevariables = male depress disability;
MODEL:
disability on depress male;
```

No distribution for the incomplete predictor!

Mplus 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:  37
*** 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:  5
   3 WARNING(S) FOUND IN THE INPUT INSTRUCTIONS

SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                         233

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

Mplus Maximum Likelihood Script

```
DATA:
file = pain.dat;
VARIABLE:
names = id male age edugroup workhrs exercise pain anxiety
  stress control interfere depress disability;
missing = all(999);
usevariables = male depress disability;
```

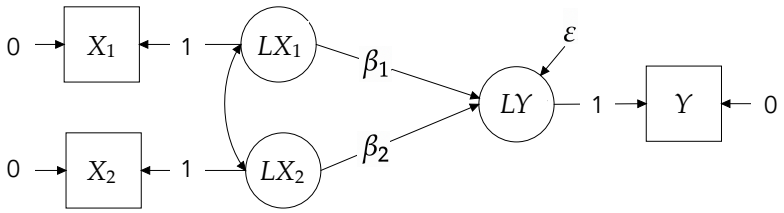

Mplus Maximum Likelihood Script

```
MODEL:
depress with male;
disability on depress (b1)
  male (b2);
OUTPUT:
stdyx;
MODEL TEST:
b1 = 0; b1 = b2;
```

Normal distribution for all predictors (even gender!)

Underlying Path Model

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



Mplus Output

```
SUMMARY OF ANALYSIS

Number of groups                                 1
Number of observations                         275

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

Mplus Output

This ominous message occurs when specifying a normal distribution for a binary variable

```
THE MODEL ESTIMATION TERMINATED NORMALLY

     THE STANDARD ERRORS OF THE MODEL PARAMETER ESTIMATES MAY NOT BE
     TRUSTWORTHY FOR SOME PARAMETERS DUE TO A NON-POSITIVE DEFINITE
     FIRST-ORDER DERIVATIVE PRODUCT MATRIX.  THIS MAY BE DUE TO THE STARTING
     VALUES BUT MAY ALSO BE AN INDICATION OF MODEL NONIDENTIFICATION.  THE
     CONDITION NUMBER IS       0.124D-11.  PROBLEM INVOLVING THE FOLLOWING
PARAMETER:
     Parameter 9, DEPRESS
```


Mplus Output

```
MODEL FIT INFORMATION

Number of Free Parameters                        9

Loglikelihood

          H0 Value                       -1707.674
          H1 Value                       -1707.674

...

Wald Test of Parameter Constraints

          Value                             61.890
          Degrees of Freedom                    2
          P-Value                           0.0000
```


Mplus Output

```
MODEL RESULTS

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

 DISABILI ON
    DEPRESS        0.318      0.040      7.867      0.000
    MALE          -0.384      0.489     -0.785      0.433

 DEPRESS  WITH
    MALE           0.270      0.193      1.394      0.163

 Means
    MALE           0.396      0.029     13.438      0.000
    DEPRESS       14.730      0.397     37.118      0.000

 Intercepts
    DISABILITY    12.253      0.653     18.776      0.000

 Variances
    MALE           0.239      0.020     11.726      0.000
    DEPRESS       38.447      3.542     10.856      0.000

 Residual Variances
    DISABILITY    14.866      1.330     11.181      0.000
```


Mplus Output

```
STANDARDIZED MODEL RESULTS

STDYX Standardization

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

 DISABILI ON
    DEPRESS        0.455      0.053      8.627      0.000
    MALE          -0.043      0.055     -0.785      0.432

...

 Residual Variances
    DISABILITY     0.794      0.048     16.699      0.000

R-SQUARE

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

    DISABILI       0.206      0.048      4.319      0.000
```

R Maximum Likelihood Script

```
# load lavaan sem package
library(lavaan)

# read data and assign variable names
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
paindata <- read.table(paste0(getwd(),"/pain.dat"))
names(paindata) <- c("id", "male", "age", "edugroup", "workhrs", "exercise", "pain",
  "anxiety", "stress", "control", "interfere", "depress", "disability")

# missing data code
paindata[paindata == 999] <- NA
```


R Maximum Likelihood Script

```
# specify model and label parameters for wald test
model <- '
disability ~ b1*depress + b2*male
depress ~~ male
'
# fit model and summarize estimates
analysis <- sem(model, paindata, missing = 'fiml', fixed.x = F,  meanstructure = T,
  mimic = "Mplus")
summary(analysis, fit.measures = TRUE, rsquare = T, standardize = T)

# wald significance test (omnibus test of all predictors)
lavTestWald(analysis, 'b1 == 0
        b2 == 0')
```


R Output

```
  Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                          9

  Number of observations                           275
  Number of missing patterns                         4

...

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -1707.674
  Loglikelihood unrestricted model (H1)      -1707.674
```


R Output

```
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  disability ~
    depress    (b1)    0.318    0.040    7.867    0.000    0.318    0.455
    male       (b2)   -0.384    0.489   -0.785    0.433   -0.384   -0.043

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  depress ~~
    male              0.270    0.193    1.394    0.163    0.270    0.089

Intercepts:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   .disability       12.253    0.653   18.776    0.000   12.253    2.833
    depress          14.730    0.397   37.118    0.000   14.730    2.376
    male              0.396    0.029   13.438    0.000    0.396    0.810
```

R Output

```
Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   .disability       14.866    1.330   11.181    0.000   14.866    0.794
    depress          38.447    3.542   10.856    0.000   38.447    1.000
    male              0.239    0.020   11.726    0.000    0.239    1.000

R-Square:
                   Estimate
    disability        0.206
```


R Output

```
$stat
[1] 61.8889

$df
[1] 2

$p.value
[1] 3.641532e-14

$se
[1] "standard"
```


Auxiliary Variables Revisited

An auxiliary variable is an ancillary variable that correlates with missingness or the analysis variables

Introducing auxiliary variables into imputation can improve power or reduce bias

The benefit of an auxiliary variable depends on the pattern and magnitude of its correlations with the analysis variables and missing data indicators

Indicator And Auxiliary Variable Correlations

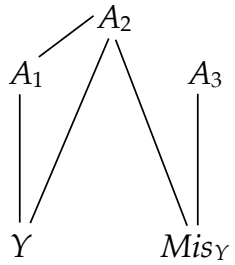
	Depression Missing Indicator	Depression
DISABILITY	0.08	0.45
MALE	-0.14	0.07
AGE	-0.08	-0.19
ANXIETY	0.03	**0.56**
STRESS	-0.04	**0.51**
CONTROL	-0.08	**-0.35**
INTERFERE	-0.09	**0.33**

Hierarchy Of Auxiliary Variables

Conditioning on A_1 can improve power but ignoring it does not introduce bias

Ignoring A_2 induces an NMAR mechanism and nonresponse bias

A_3 has no effect on bias and power and should not be used



Practical Conclusions

There are no A_2 variables with strong enough correlations to introduce bias if ignored

Four A_1 variables are moderately or strongly correlated with depression (the focal predictor)

Estimating the model with additional variables can reduce standard errors and improve power

Model Specification Rules

Correlate each auxiliary variable with …

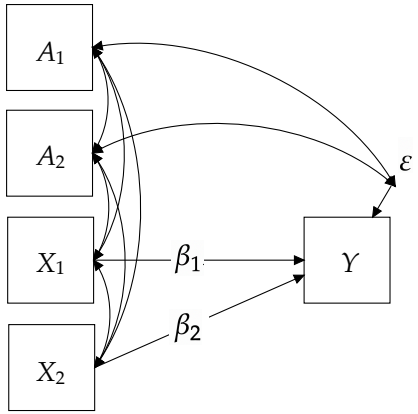
Manifest predictor variables

Other auxiliary variables

The residual terms of all dependent variables

Auxiliary variables never correlate with latent variables

Saturated Correlates Path Diagram



Summary Of Multiple Imputation Estimates

Results based on all available data from 250 participants

Parameter	Est.	*SE*	*z*	*p*
Intercept	12.25	0.65	18.96	< .001
DEPRESS slope	0.32	0.04	7.97	< .001
MALE slope	-0.39	0.49	-0.80	0.425

Interpretations

Auxiliary variables don't affect the interpretations

Controlling for gender, a one-unit increase in depression predicts an increase in psychosocial disability of .32 ($p < .001$)

Controlling for depression, males have a psychosocial disability mean that is .39 points lower than that of females ($p = .43$)

Mplus Maximum Likelihood Script

```
DATA:
file = pain.dat;
VARIABLE:
names = id male age edugroup workhrs exercise pain anxiety
  stress control interfere depress disability;
missing = all(999);
usevariables = male depress disability;
auxiliary = (m) anxiety control;
```

← Missing data auxiliary variables

Mplus Maximum Likelihood Script

```
MODEL:
depress with male;
disability on depress (b1)
  male (b2);
OUTPUT:
stdyx;
MODEL TEST:
b1 = 0; b1 = b2;
```

← Normal distribution for all predictors (even gender!)

Mplus Output

```
SUMMARY OF ANALYSIS

Number of groups                                 1
Number of observations                         275

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

Observed dependent variables

  Continuous
   DISABILITY

Observed independent variables
   MALE        DEPRESS

Observed auxiliary variables
   ANXIETY     CONTROL
```

81

Mplus Output

```
MODEL FIT INFORMATION

Number of Free Parameters                        9

Loglikelihood Including the Auxiliary Part

          H0 Value                       -3251.124
          H1 Value                       -3251.124

...

Wald Test of Parameter Constraints

          Value                             63.587
          Degrees of Freedom                    2
          P-Value                           0.0000
```

82

Mplus Output

```
MODEL RESULTS

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

 DISABILI ON
    DEPRESS        0.317      0.040      7.974      0.000
    MALE          -0.389      0.487     -0.799      0.425

 DEPRESS  WITH
    MALE           0.244      0.193      1.263      0.207

 Means
    MALE           0.396      0.029     13.438      0.000
    DEPRESS       14.855      0.396     37.539      0.000

 Intercepts
    DISABILITY    12.251      0.646     18.955      0.000

 Variances
    MALE           0.239      0.020     11.726      0.000
    DEPRESS       39.317      3.637     10.811      0.000

 Residual Variances
    DISABILITY    14.876      1.321     11.258      0.000
```

83

Mplus Output

```
STANDARDIZED MODEL RESULTS

STDYX Standardization

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

 DISABILI ON
    DEPRESS        0.458      0.052      8.878      0.000
    MALE          -0.044      0.055     -0.799      0.424

...

 Residual Variances
    DISABILITY     0.791      0.047     16.864      0.000

R-SQUARE

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

    DISABILI       0.209      0.047      4.443      0.000
```

84

R Maximum Likelihood Script

```
# load lavaan sem package with semtools
library(lavaan)
library(semTools)

# read data and assign variable names
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
paindata <- read.table(paste0(getwd(),"/pain.dat"))
names(paindata) <- c("id", "male", "age", "edugroup", "workhrs", "exercise", "pain",
  "anxiety", "stress", "control", "interfere", "depress", "disability")

# missing data code
paindata[paindata == 999] <- NA

# auxiliary variables
auxvars <- c("anxiety", "control")
```


R Maximum Likelihood Script

```
# specify model and label parameters for wald test
model <- '
disability ~ b1*depress + b2*male
depress ~~ male
'
# fit model and summarize estimates
analysis <- sem.auxiliary(model = model, aux = auxvars, data = paindata, missing = 'fiml',
  fixed.x = F, meanstructure = T, mimic = "Mplus")
summary(analysis, fit.measures = T, rsquare = T, standardize = T)

# wald significance test (omnibus test of all predictors)
lavTestWald(analysis, 'b1 == 0
        b2 == 0')
```


R Output

```
  Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                         20

  Number of observations                           275
  Number of missing patterns                         6

...

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -3251.124
  Loglikelihood unrestricted model (H1)      -3251.124
```


R Output

```
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  disability ~
    depress    (b1)   0.317    0.040    7.974    0.000    0.317    0.458
    male       (b2)  -0.389    0.487   -0.799    0.425   -0.389   -0.044

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  depress ~~
    male              0.244    0.193    1.263    0.207    0.244    0.079
  anxiety ~~
    control          -7.256    1.553   -4.674    0.000   -7.256   -0.300
   .disability        1.488    0.974    1.527    0.127    1.488    0.084
  control ~~
   .disability       -3.887    1.209   -3.214    0.001   -3.887   -0.192
  anxiety ~~
    depress          16.327    2.135    7.648    0.000   16.327    0.564
  control ~~
    depress         -11.692    2.181   -5.360    0.000  -11.692   -0.356
  anxiety ~~
    male              0.190    0.139    1.361    0.174    0.190    0.084
  control ~~
    male             -0.150    0.155   -0.968    0.333   -0.150   -0.058
```

R Output

```
Intercepts:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   .disability       12.251    0.646   18.955    0.000   12.251    2.826
    depress          14.855    0.396   37.539    0.000   14.855    2.369
    male              0.396    0.029   13.438    0.000    0.396    0.810
    anxiety          11.613    0.284   40.935    0.000   11.613    2.517
    control          20.764    0.316   65.673    0.000   20.764    3.960

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   .disability       14.876    1.321   11.258    0.000   14.876    0.791
    depress          39.317    3.637   10.811    0.000   39.317    1.000
    male              0.239    0.020   11.726    0.000    0.239    1.000
    anxiety          21.290    1.853   11.489    0.000   21.290    1.000
    control          27.490    2.344   11.726    0.000   27.490    1.000

R-Square:
                   Estimate
    disability        0.209
```