1. Missing Data Mechanisms

Preliminary Ideas

Missing values exist in principle, but are not observed in our sample

The hypothetically-complete data, $data_{(com)}$, can be partitioned into two parts

$data_{(obs)}$ = observed data, and $data_{(mis)}$ = unseen missing scores (i.e., "latent" variables)

Partitioning The Data

Hypothetically complete data

Y_1	Y_2	Y_3
4	4	3
3	3	5
7	1	6
2	1	6
5	9	3
3	2	2
1	6	7
9	4	9
2	5	6

$data_{(com)}$

Observed data

Y_1	Y_2	Y_3
4	4	3
3	NA	5
7	1	6
NA	1	6
5	9	3
3	NA	NA
1	6	7
9	4	9
2	NA	6

$data_{(obs)}$

Missing data

Y_1	Y_2	Y_3
	3	
2		
	2	2
	5	

$data_{(mis)}$

Patterns Versus Mechanisms

A missing data pattern describes the location of the holes in the data, but says nothing about why the data are missing

The missing data mechanism describes possible associations between the data and missingness

Is missingness related to $data_{(obs)}$ or $data_{(mis)}$?

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

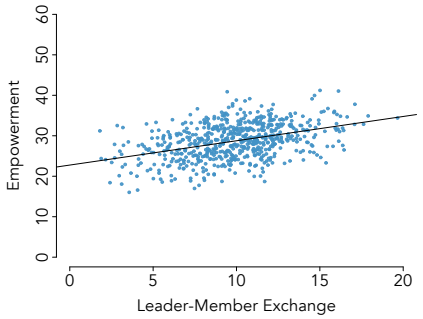
lmxquality.dat

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

Artificial Data Example

Measures of employee empowerment and employee-supervisor relationship quality (leader-member exchange)

50% of relationship scores are missing under an MCAR, MAR, or NMAR mechanism



Mechanisms Implemented

MCAR ➠ missingness is unrelated to both SES and math achievement

MAR ➠ achievement scores are missing as SES decreases (e.g., low-SES students are more mobile)

NMAR ➠ achievement scores are missing as math achievement decreases (e.g., low-achieving students fail to complete the exam)

Not Missing At Random Mechanism

An NMAR mechanism is one where the propensity for a missing value is related to the observed or missing (latent) parts of the data

$$p\left(missing \mid data_{(obs)}, data_{(mis)}\right)$$

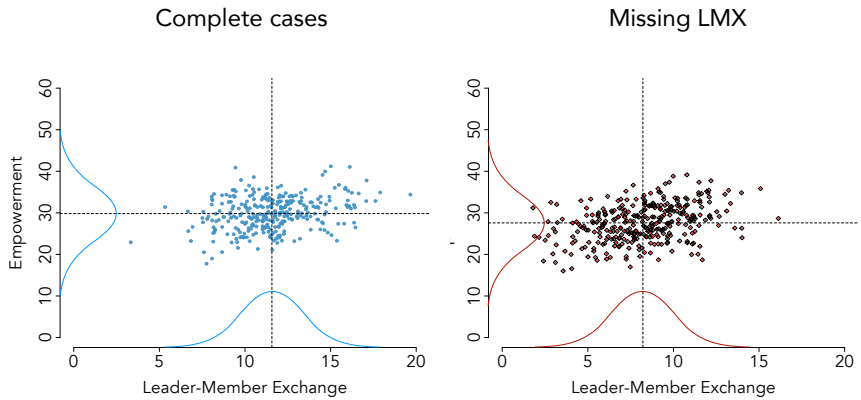
Missingness depends on the unseen latent scores

Hypothetical Substantive Example

The likelihood of a missing relationship quality (leader-member) score increases as relationship quality decreases

Employees who have poor relationships with their supervisors are less likely to report their relationship quality (e.g., out of concern for making the situation worse)

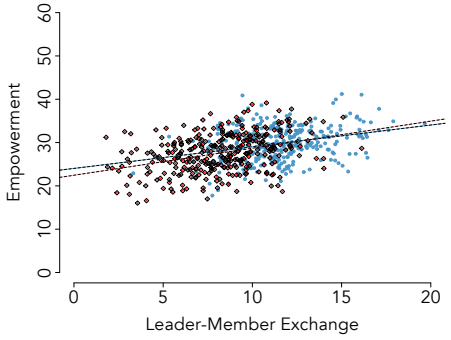
NMAR Comparison



Implications Of NMAR

Complete and incomplete cases have different model parameters

The data alone do not contain enough information to estimate both models



Simplification: Missing At Random

MAR posits that the probability of a missing value is related only to the observed part of the data

$$p\left(missing \mid data_{(obs)}, data_{(mis)}\right) = p\left(missing \mid data_{(obs)}\right)$$

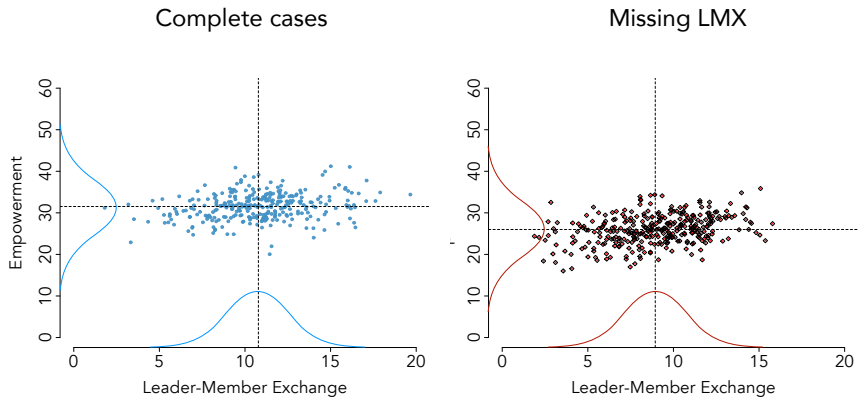
The unseen latent scores carry no additional information above and beyond the observed data

Hypothetical Substantive Example

The likelihood of a missing relationship quality (leader-member) score increases as employee empowerment decreases

Employees who do not feel empowered are less likely to report their relationship quality (e.g., because they are apathetic about their job)

MAR Comparison

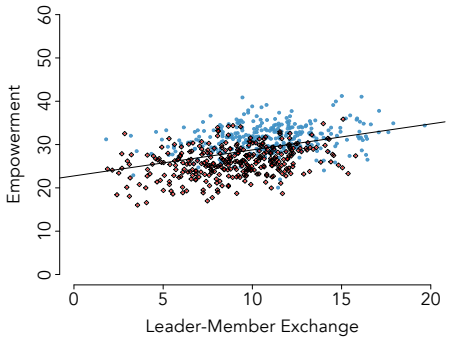


Implications Of MAR

Complete and incomplete cases share the same model

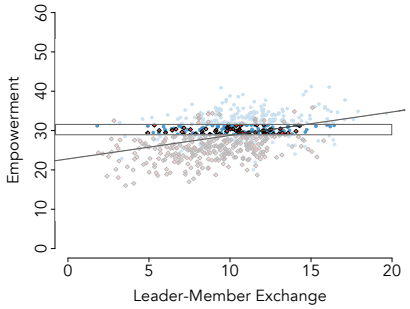
Using the analysis model to generate imputations gives accurate estimates

e.g., Maximum likelihood, Bayes, multiple imputation

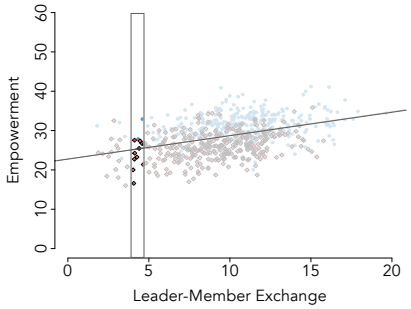


Imputation Preview

If Y is observed, missing X values fall along a horizontal slice of the distribution



If X is observed, missing Y values fall along a vertical slice of the distribution



Further Simplification: Missing Completely At Random

An MCAR mechanism is one where missingness is completely unrelated to the data

$$p\left(missing \mid data_{(obs)}, data_{(mis)}\right) = p(missing)$$

All participants have the same chance of missing data, regardless of their characteristics

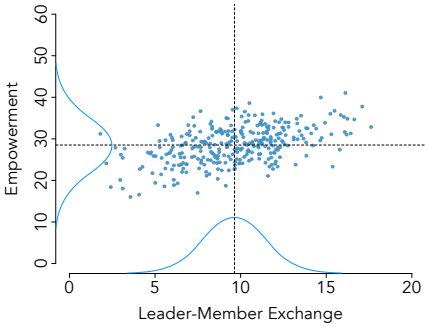
Hypothetical Substantive Example

The likelihood of a missing relationship quality (leader-member) score is unrelated to empowerment and relationship quality

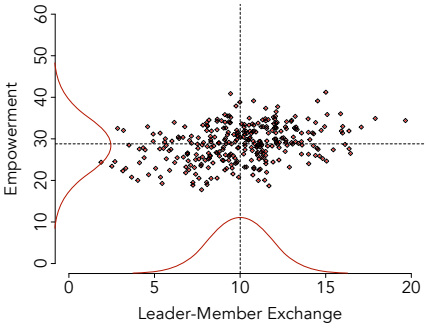
Researchers use a planned missing data design where the relationship quality measure is administered to only 50% of the respondents

MCAR Comparison

Complete cases



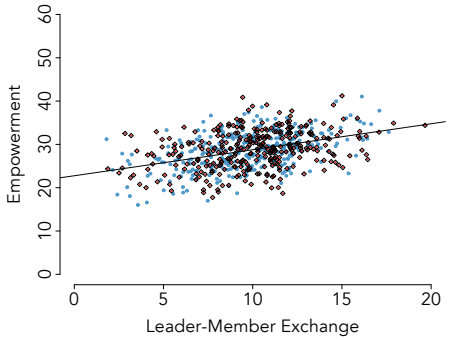
Missing LMX



Implications Of MCAR

Cases with missing data are the same as those with complete data, on average

Excluding missing cases yields valid estimates, albeit with lower precision



21

Why Mechanisms Matter

Mechanisms function as assumptions, estimates are biased when assumptions do not hold

Some older approaches require MCAR (others make no attempt to satisfy any mechanism)

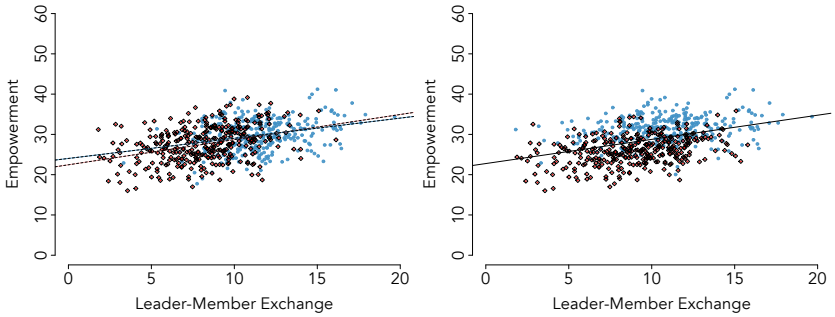
Maximum likelihood, Bayesian estimation, and multiple imputation assume MAR (or MCAR)

22

NMAR vs. MAR

NMAR is the true process

MAR analyses are misspecified and exhibit biases

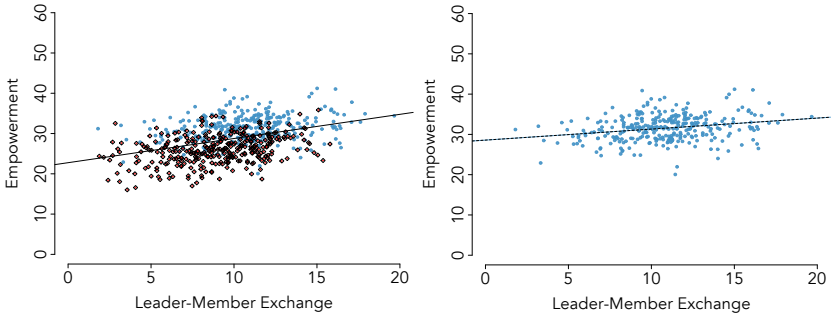


23

MAR vs. MCAR

MAR is the true process

MCAR analyses are misspecified and exhibit biases



24

Diagnosing Mechanisms

MCAR is the only mechanism with testable propositions (e.g, no predictors of missingness)

Create a missing data indicator (e.g., 0 = complete, 1 = missing) for each incomplete variable and search for its correlates

This strategy can rule out MCAR but says nothing about the plausibility of MAR or NMAR

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

The MAR assumption must be considered on a analysis-by-analysis basis

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

MAR means that leader-member exchange is the only correlate of the the empowerment missing data indicator (and vice versa)

Indicator Correlations

	Empowerment Missing Indicator	LMX Missing Indicator
TURNOVER	-0.06	0.02
MALE	-0.01	0.05
EMPOWER	NA	0.03
LMXQUALITY	**0.38**	NA
JOBSAT	**0.19**	-0.05
CLIMATE	0.04	-0.02
ORGSIZE	-0.06	0.00

Practical Conclusions

Correlates of missingness rule out MCAR but provide no evidence about MAR vs. NMAR

Job satisfaction and leader-member exchange correlate with the empowerment indicator

MAR estimation methods automatically adjust for other variables in the analysis, but job satisfaction is not in the model

Auxiliary Variables

An auxiliary variable is outside of the analysis model but correlates with missingness or the analysis variables

Introducing auxiliary variables into the missing data handling procedure may improve power or reduce bias

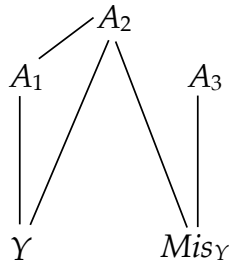
The benefit of an auxiliary variable depends on the pattern and magnitude of its correlations with the analysis variables and missing data indicators

Hierarchy Of Auxiliary Variables

Conditioning on A_1 improves power but ignoring it does not introduce bias

Ignoring A_2 induces an NMAR mechanism and nonresponse bias

A_3 cannot introduce bias nor can it increase power



Example

Job satisfaction correlates with the empowerment missingness indicator

Satisfaction also correlates with empowerment, r = .29, and leader-member exchange, r = .42

Job satisfaction parallels A_2 in the path diagram

Practical Recommendations

MAR-based methods (e.g., maximum likelihood, Bayesian, multiple imputation) may not be a perfect solution, even with auxiliary variables

The observed data cannot differentiate MAR and NMAR mechanisms, expert judgment is required

NMAR-based procedures are difficult to implement, so it is convenient to assume MAR

Mplus Indicator Correlation Script

```
DATA:
file = lmxquality.dat;
VARIABLE:
names = employee team turnover male empower lmxquality jobsat climate orgsize;
usevariables = turnover - orgsize emp_ind lmx_ind;
missing = all(999);
DATA MISSING:
names = empower lmxquality;
type = missing;
binary = emp_ind lmx_ind;
MODEL:
emp_ind lmx_ind with turnover - orgsize;
turnover - orgsize with turnover - orgsize;
emp_ind with empower@0;
lmx_ind with lmxquality@0;
OUTPUT:
stdyx;
```


Mplus Output

```
STANDARDIZED MODEL RESULTS

STDYX Standardization

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

 EMP_IND  WITH
    TURNOVER         -0.029      0.045     -0.639      0.523
    MALE             -0.007      0.040     -0.175      0.861
    EMPOWER           0.000      0.000    999.000    999.000
    LMXQUALITY        0.380      0.034     11.076      0.000
    JOBSAT            0.165      0.043      3.826      0.000
    CLIMATE           0.042      0.041      1.023      0.306
    ORGSIZE          -0.061      0.042     -1.459      0.145

 LMX_IND  WITH
    TURNOVER          0.009      0.040      0.229      0.819
    MALE              0.054      0.040      1.350      0.177
    EMPOWER           0.029      0.040      0.728      0.467
    LMXQUALITY        0.000      0.000    999.000    999.000
    JOBSAT           -0.032      0.039     -0.814      0.416
    CLIMATE          -0.011      0.043     -0.254      0.800
    ORGSIZE          -0.003      0.041     -0.070      0.945
```


Mplus Output

```
STANDARDIZED MODEL RESULTS

STDYX Standardization

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

 EMPOWER  WITH
    LMXQUALITY        0.367      0.034     10.693      0.000
    JOBSAT            0.284      0.040      7.164      0.000
    CLIMATE           0.209      0.043      4.880      0.000
    ORGSIZE          -0.062      0.044     -1.415      0.157

 LMXQUALI WITH
    JOBSAT            0.418      0.034     12.225      0.000
    CLIMATE           0.029      0.042      0.680      0.497
    ORGSIZE          -0.061      0.042     -1.453      0.146
```

SPSS Indicator and Correlation Script

```
* set working directory.
CD "YOUR-FILE-PATH".

* read data and compute missing data indicators.
DATA LIST free file = "lmxquality.dat"
  /employee team turnover male empower lmxquality jobsat climate orgsize.
MISSING VALUES all (999).
RECODE empower lmxquality (999=1)(else=0) empower_ind lmxquality_ind.

* indicator correlations.
CORRELATIONS ind_empower ind_lmxquality with  turnoverint to orgsize.
* analysis variables and auxiliary variable correlations.
CORRELATIONS empower lmxquality with  jobsat.
```


SPSS Output

Correlations

		turnover	male	empower	lmxquality	jobsat	climate	orgsize
empower_ind	Pearson Correlation	-.061	-.012	.[a]	.384	.194	.040	-.056
	Sig. (2-tailed)	.138	.765	.000	.000	.000	.340	.176
	N	598	630	528	604	600	570	594
lmxquality_ind	Pearson Correlation	.015	.054	.032	.[a]	-.049	-.018	.001
	Sig. (2-tailed)	.722	.172	.469	.000	.229	.675	.977
	N	598	630	528	604	600	570	594

a. Cannot be computed because at least one of the variables is constant.

Correlations

		jobsat
empower	Pearson Correlation	.287
	Sig. (2-tailed)	.000
	N	523
lmxquality	Pearson Correlation	.423
	Sig. (2-tailed)	.000
	N	574

R Indicator Correlation Script

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

# missing data indicators
data[data == 999] <- NA
data$lmxquality_ind <- as.numeric(is.na(data$lmxquality))
data$empower_ind <- as.numeric(is.na(data$empower))
```


R Indicator Correlation Script

```
# indicator correlations
cor(data[c("lmxquality_ind")], data[c("turnover", "male",
  "empower","jobsat","climate","orgsize")], use = "pairwise.complete.obs")
cor(data[c("empower_ind")], data[c("turnover", "male",
  "lmxquality","jobsat","climate","orgsize")], use = "pairwise.complete.obs")

# correlations between analysis variables and auxiliary variable
cor(data[c("jobsat")], data[c("empower","lmxquality")],
  use = "pairwise.complete.obs")
```

R Output

```
                turnover       male    empower      jobsat     climate    orgsize
lmxquality_ind 0.01459054 0.05449313 0.03160615 -0.04922508 -0.01759351 0.00118045

              turnover        male lmxquality    jobsat    climate     orgsize
empower_ind -0.06066996 -0.01190723  0.3843457 0.1935601 0.04006408 -0.05562753

         empower lmxquality
jobsat 0.2866214  0.4230336
```


Stata Indicator Correlation Script

```
// set working directory
cd "YOUR-FILE-PATH"

// read data
clear
infile employee team turnover male empower lmxquality jobsat climate orgsize using “lmxquality.dat”

// missing data indicator
recode empower (min/998 = 0)(999 = 1), generate(emp_ind)
recode lmxquality (min/998 = 0)(999 = 1), generate(lmx_ind)

// recode missing data in original data
recode turnover - orgsize (999 = .)

// indicator correlations
pwcorr emp_ind lmx_ind turnover - orgsize
```


Stata Output

```
             |  emp_ind  lmx_ind turnover     male  empower lmxqua~y   jobsat
-------------+------------------------------------------------------------------
     emp_ind |   1.0000
     lmx_ind |  -0.0912   1.0000
    turnover |  -0.0607   0.0146   1.0000
        male |  -0.0119   0.0545  -0.0585   1.0000
     empower |         .   0.0316  -0.1770   0.2582   1.0000
  lmxquality |   0.3843         .  -0.2099   0.0785   0.4036   1.0000
      jobsat |   0.1936  -0.0492  -0.3128   0.1381   0.2866   0.4230   1.0000
     climate |   0.0401  -0.0176  -0.2412   0.0930   0.1894   0.0313   0.2665
     orgsize |  -0.0556   0.0012   0.0573   0.0182  -0.0609  -0.0547  -0.0674

             |  climate  orgsize
-------------+------------------
     climate |   1.0000
     orgsize |  -0.2261   1.0000
```