Introduction to Missing Data Analyses

2020 Summer Institute
for the UC Adolescence Consortium

Craig Enders

University of California - Los Angeles

cenders@psych.ucla.edu

Body Attitudes Data

Questionnaire data from a study of body attitudes in a sample of 500 middle school students

Variables include body mass index (BMI), five questionnaire items measuring negative body attitudes, and past history of being bullied (0 = never bullied, 1 = history of being bullied)

All questionnaire items measured on a 7-point scale

Negative Body Attitudes Questionnaire

	Strongly Disagree ⟷ Strongly Agree						
1.My hips seem to broad to me.	1	2	3	4	5	6	7
2. I have a strong desire to be thinner.	1	2	3	4	5	6	7
3. I think I'm too thick.	1	2	3	4	5	6	7
4. Some parts of my body look swollen.	1	2	3	4	5	6	7
5. My belly looks as if I'm pregnant.	1	2	3	4	5	6	7

bodyattitudes.dat

Variable	Name	Missing %	Scaling
Identifier variable	ID	0	Integer index
History of being bullied	BULLIED	10.4	0 = not bullied, 1 = bullied
Body mass index	BMI	8.0	Continuous
Attitude item 1 (hips too broad)	BATT1	13.4	7-point ordinal scale
Attitude item 2 (desire to be thinner)	BATT2	12.2	7-point ordinal scale
Attitude item 3 (too thick)	BATT3	0	7-point ordinal scale
Attitude item 4 (body looks swollen)	BATT4	12.4	7-point ordinal scale
Attitude item 5 (belly looks pregnant)	BATT5	0	7-point ordinal scale
Negative body attitudes composite	NEGBODYATT	29.6	Sum of five items

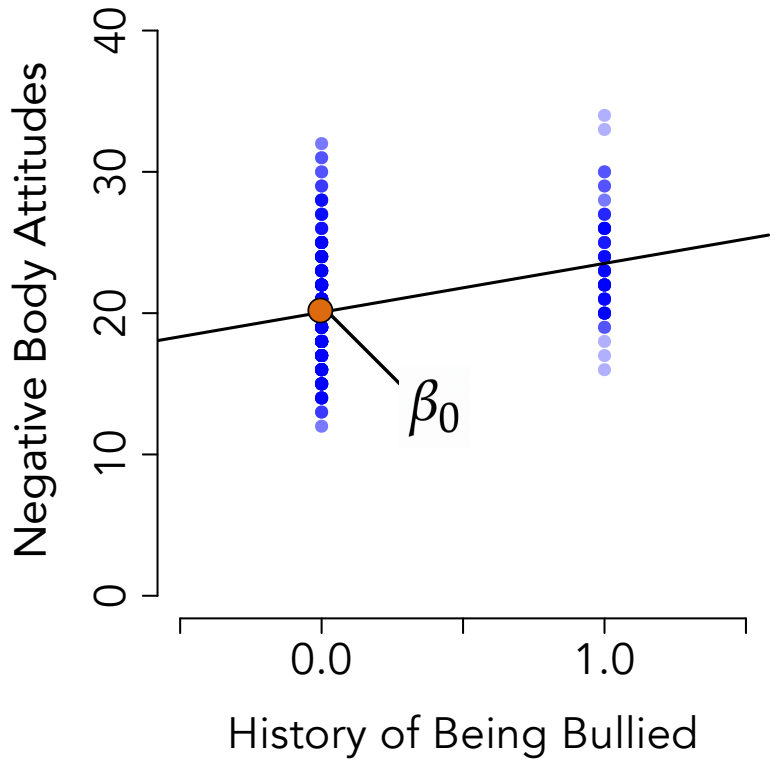
Substantive Analysis

Negative body attitudes scale score regressed on body mass index and bullied indicator

$$NEGBODYATT_i = \beta_0 + \beta_1(BMI_i) + \beta_2(BULLIED_i) + \varepsilon_i$$

Substantive goal is to compare bullied and not bullied groups while controlling for BMI

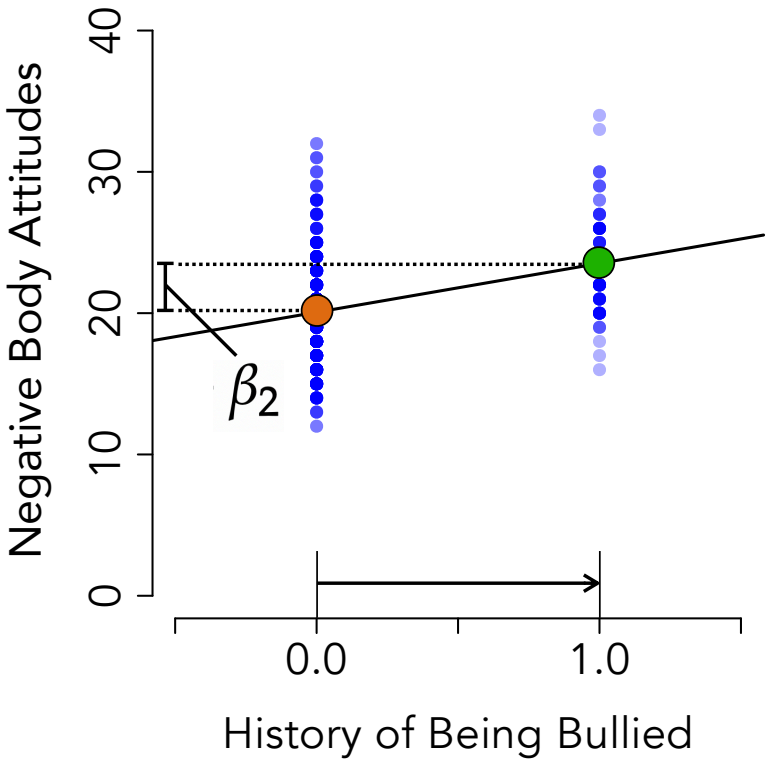
Regression Intercept



The intercept is the expected body attitudes score when bullied = 0

i.e., The mean for girls with no history of being bullied

Regression Slope



The slope is the change in body attitudes score between the not bullied to bullied group

i.e., The mean difference for girls who have been bullied

Missing Data Processes

The missing data process (mechanism) dictates the accuracy of a missing data handling method

Missing values could be haphazard and unrelated to the variables in the analysis model

Missingness is often systematic and could be related to observed data or the unseen values

Observed And Missing Data

Hypothetically complete data

Y	X_1	X_2
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

=

Observed data

Y	X_1	X_2
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

+

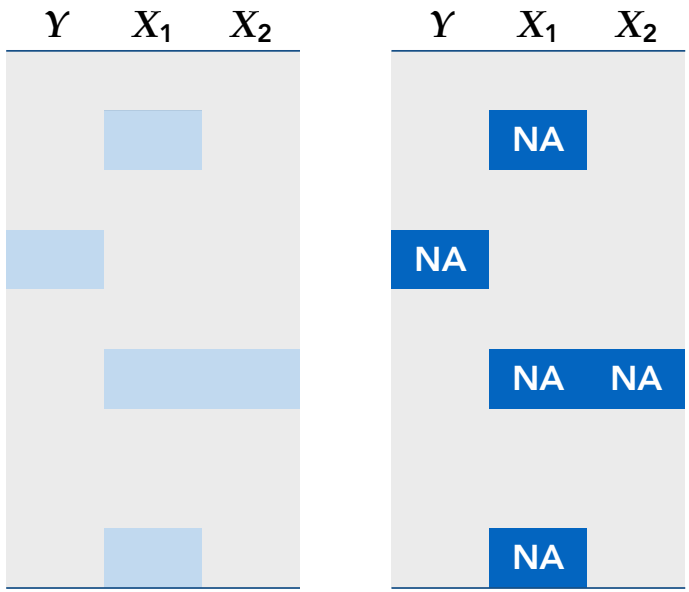
Missing data (latent, unseen)

Y	X_1	X_2
	3	
2		
	2	2
	5	

Missing Completely At Random (MCAR)

Haphazard non-response, variables in the analysis do not predict missingness

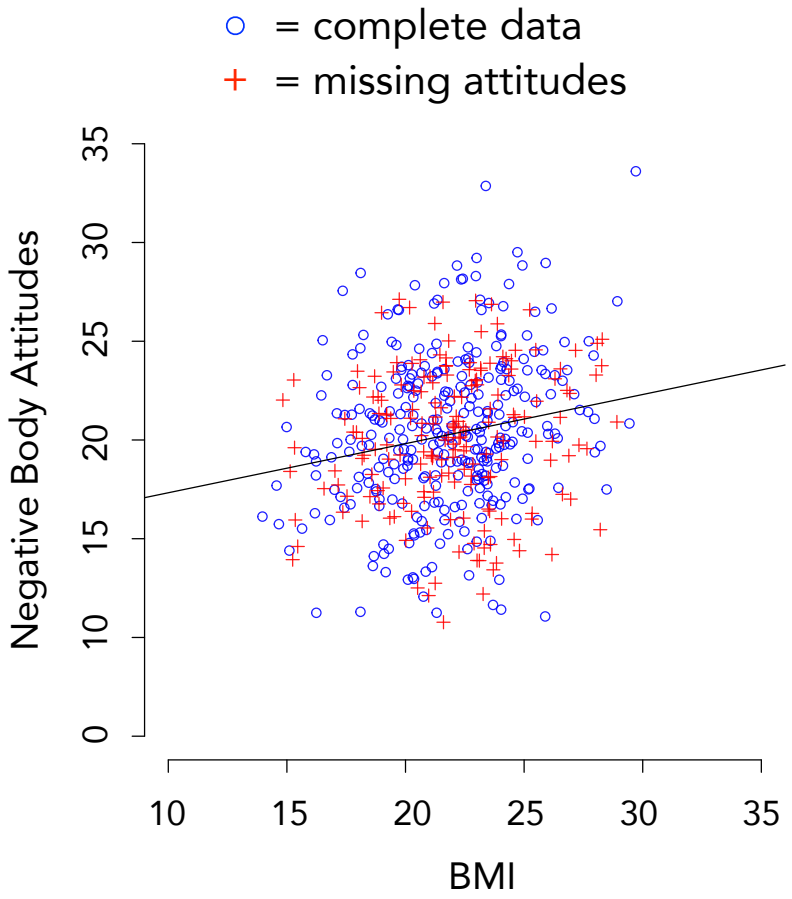
e.g., Items missing due to a purely chance process



Missing Completely At Random (MCAR)

Haphazard non-response, variables in the analysis do not predict missingness

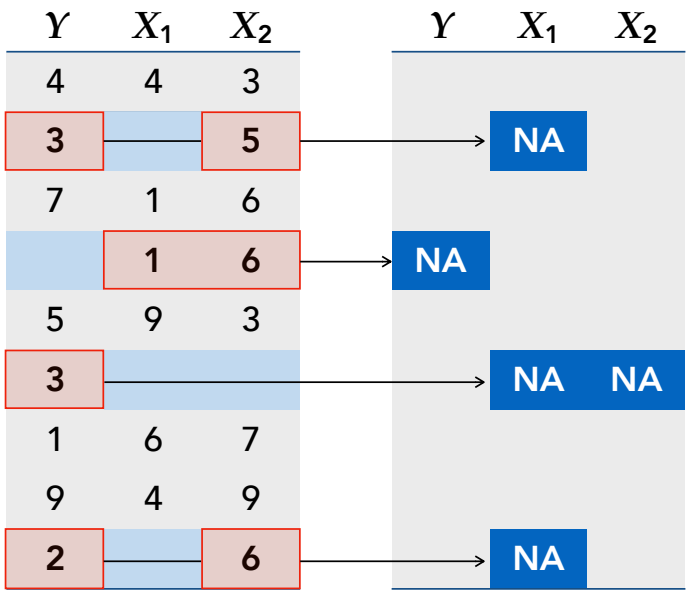
e.g., Items missing due to a purely chance process



Missing At Random (MAR)

Systematic non-response, observed scores in the analysis predict missingness

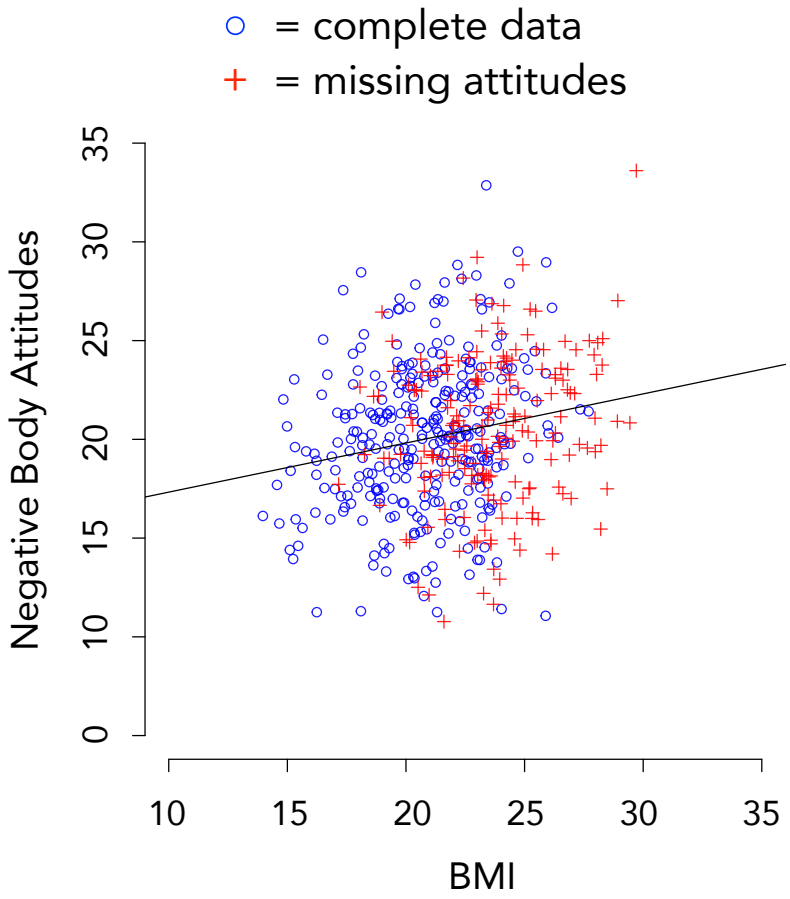
e.g., Observed variables such as BMI predict whether attitude items are missing



Missing At Random (MAR)

Systematic non-response, observed scores in the analysis predict missingness

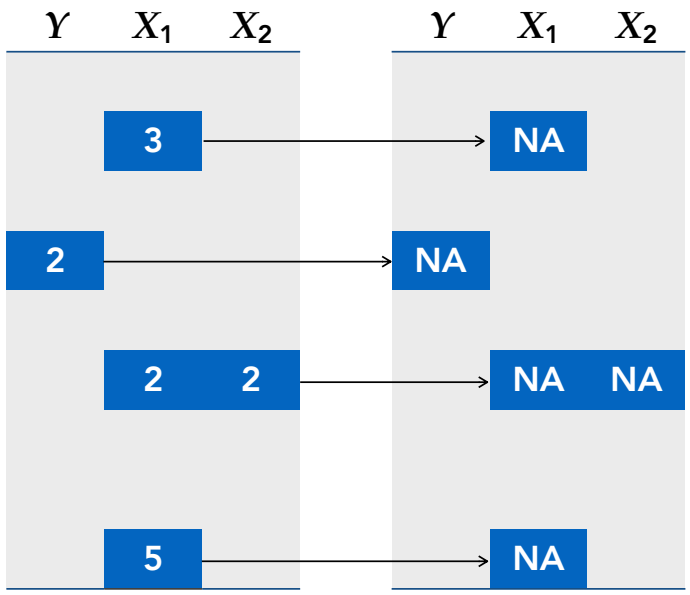
e.g., Observed variables such as BMI predict whether attitude items are missing



Not Missing At Random (NMAR)

Systematic non-response, unseen (latent) scores in the analysis predict missingness

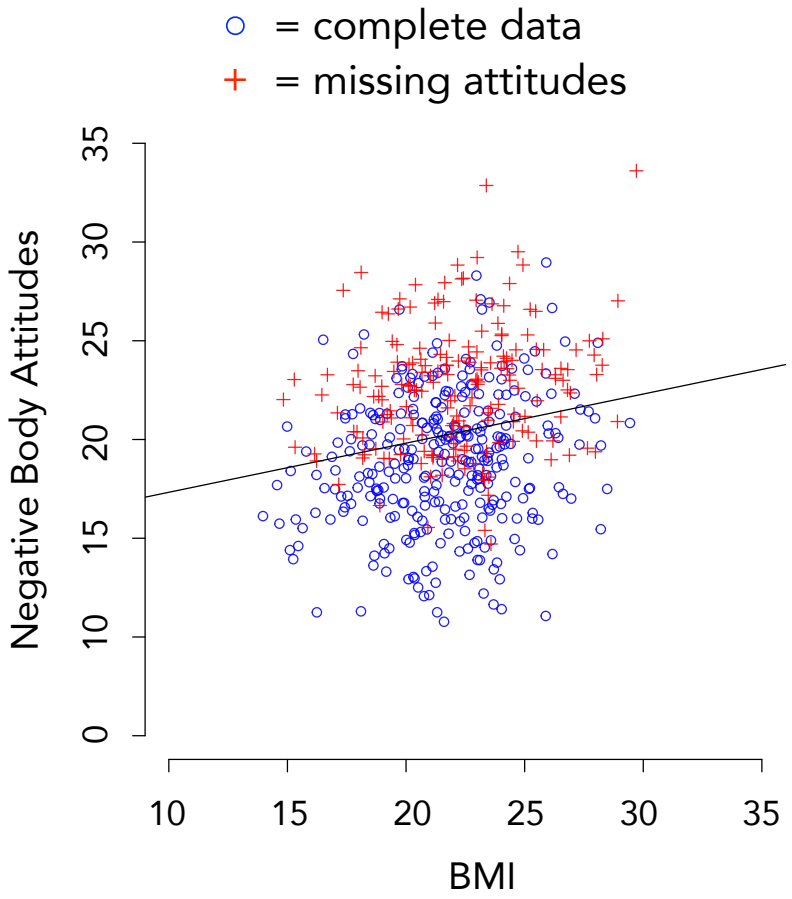
e.g., Attitude items are missing for participants with the most negative attitudes



Not Missing At Random (NMAR)

Systematic non-response, unseen (latent) scores in the analysis predict missingness

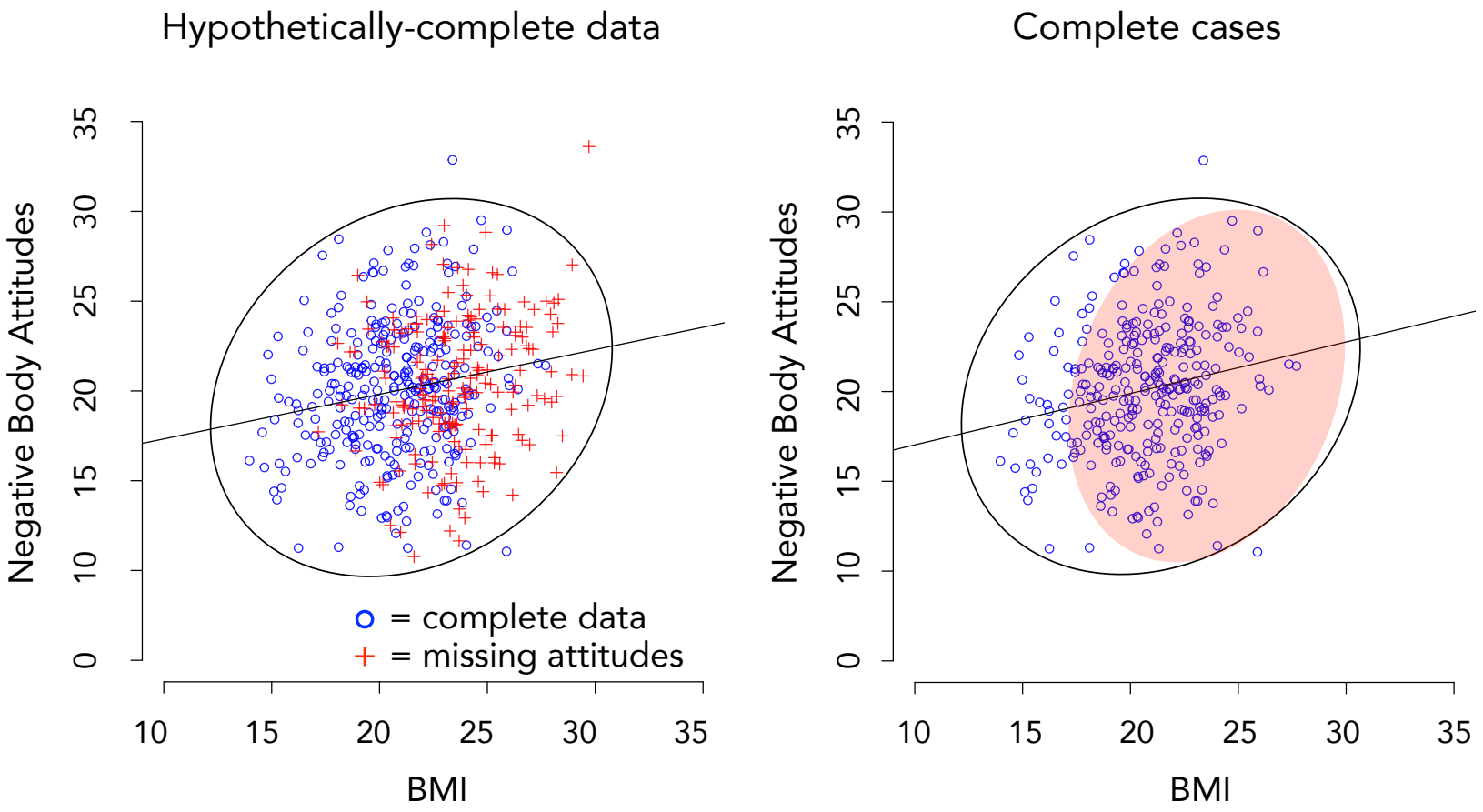
e.g., Attitude items are missing for participants with the most negative attitudes



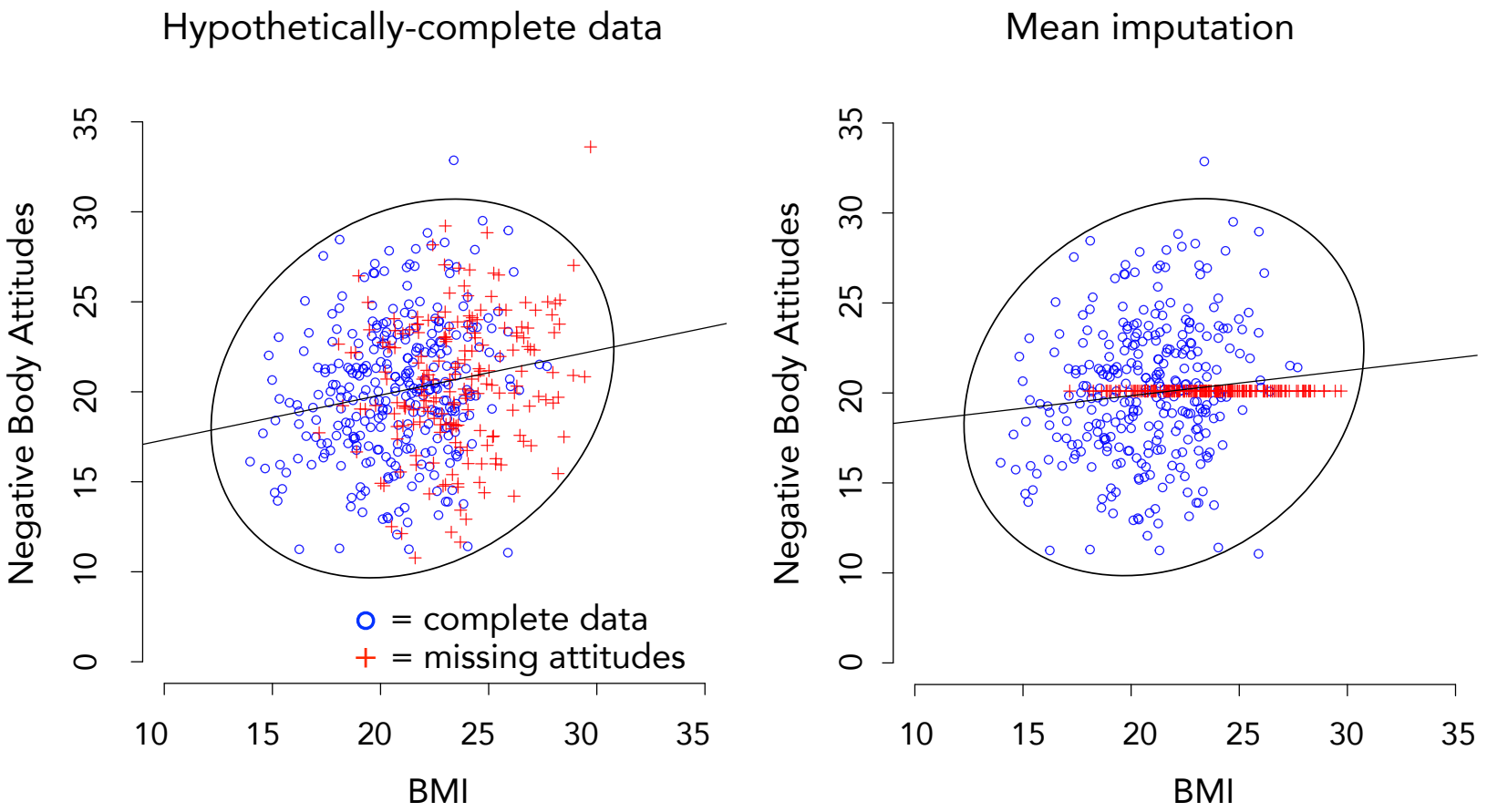
Some Popular But Flawed Methods

Method	Properties
Exclude incomplete data records (completely or analysis-by-analysis)	Decreases power, requires unsystematic missingness, biased with systematic processes
Fill in missing values with the arithmetic mean	Atheoretical, severe bias under any process
Replace missing scores with predicted values from a regression equation	Distorts measures of variation and correlation
In longitudinal data, replace missing scores with the last observation from a preceding wave	Atheoretical, biased under any process

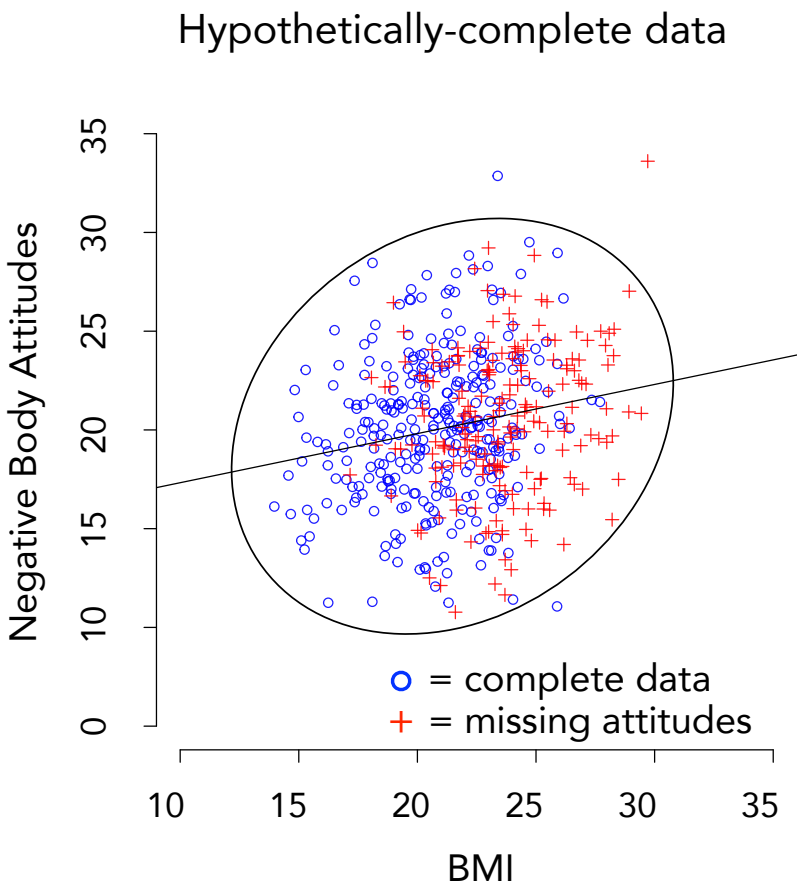
Complete-Case Analysis (Deletion)

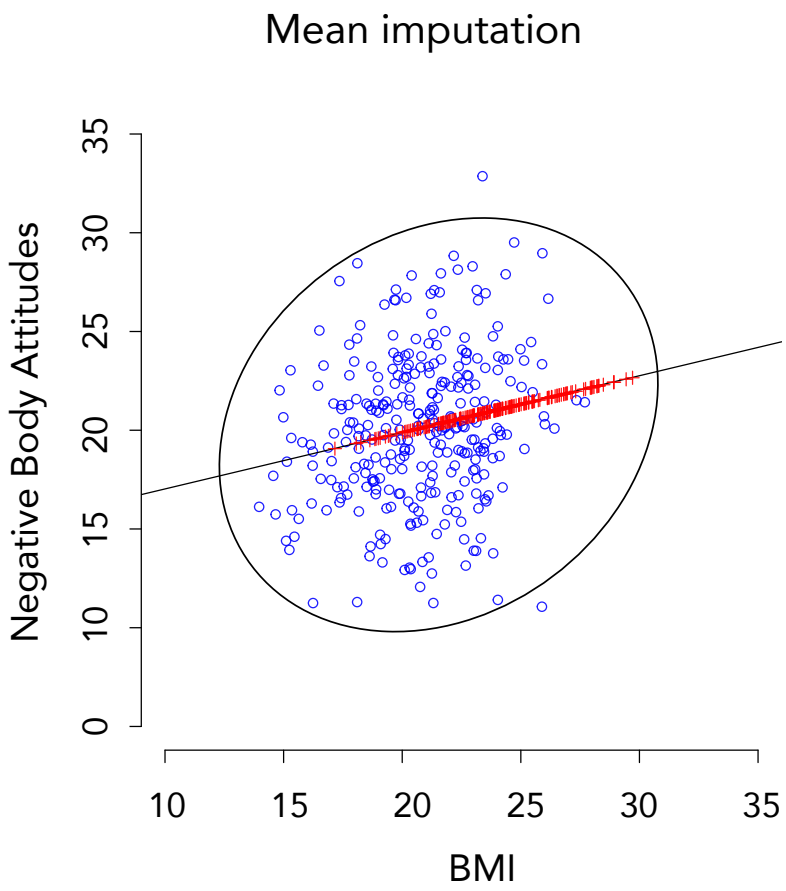


Mean Imputation



Regression Imputation





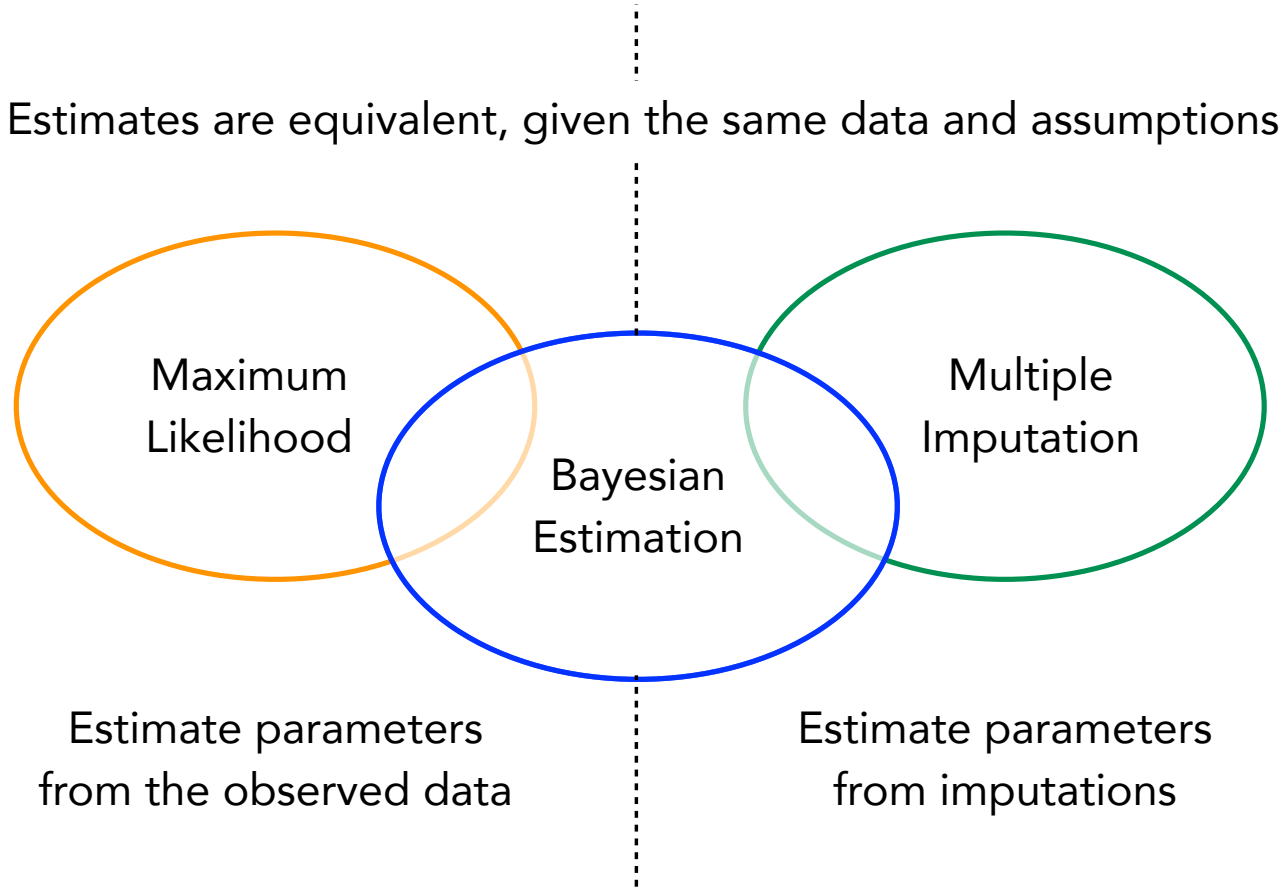
Current Recommendations

Methods that assume systematic non-response due to observed scores (MAR) are often optimal

Easy to implement, widely available in software packages, and flexible for different applications

Maximum likelihood, Bayesian estimation, and multiple imputation assume an MAR process

Relations Among Methods



Practical Benefits

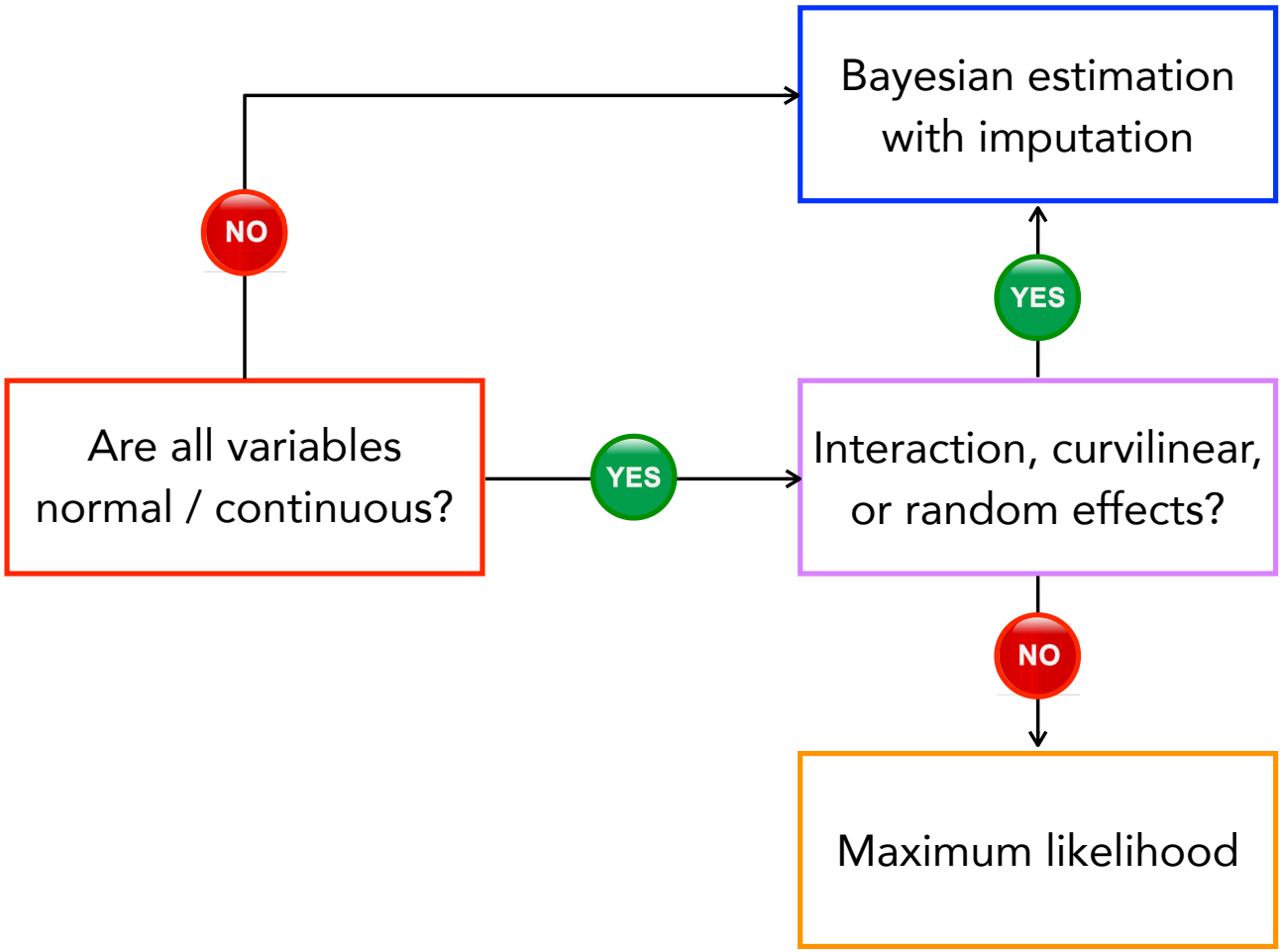
If missingness is related to observed scores …

a) Substantial missing data rates cause no bias

b) Power is maximized because all available data contribute to the analysis

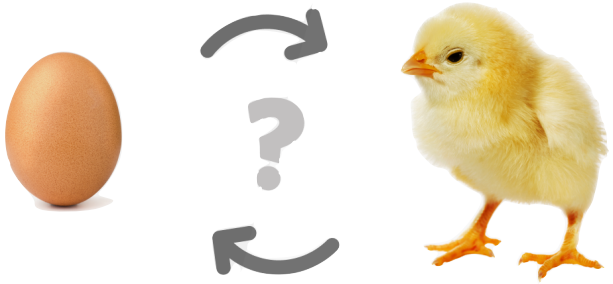
c) Performance relative to other methods (e.g., deletion) improves as missing data rates increase

Choosing A Missing Data Method



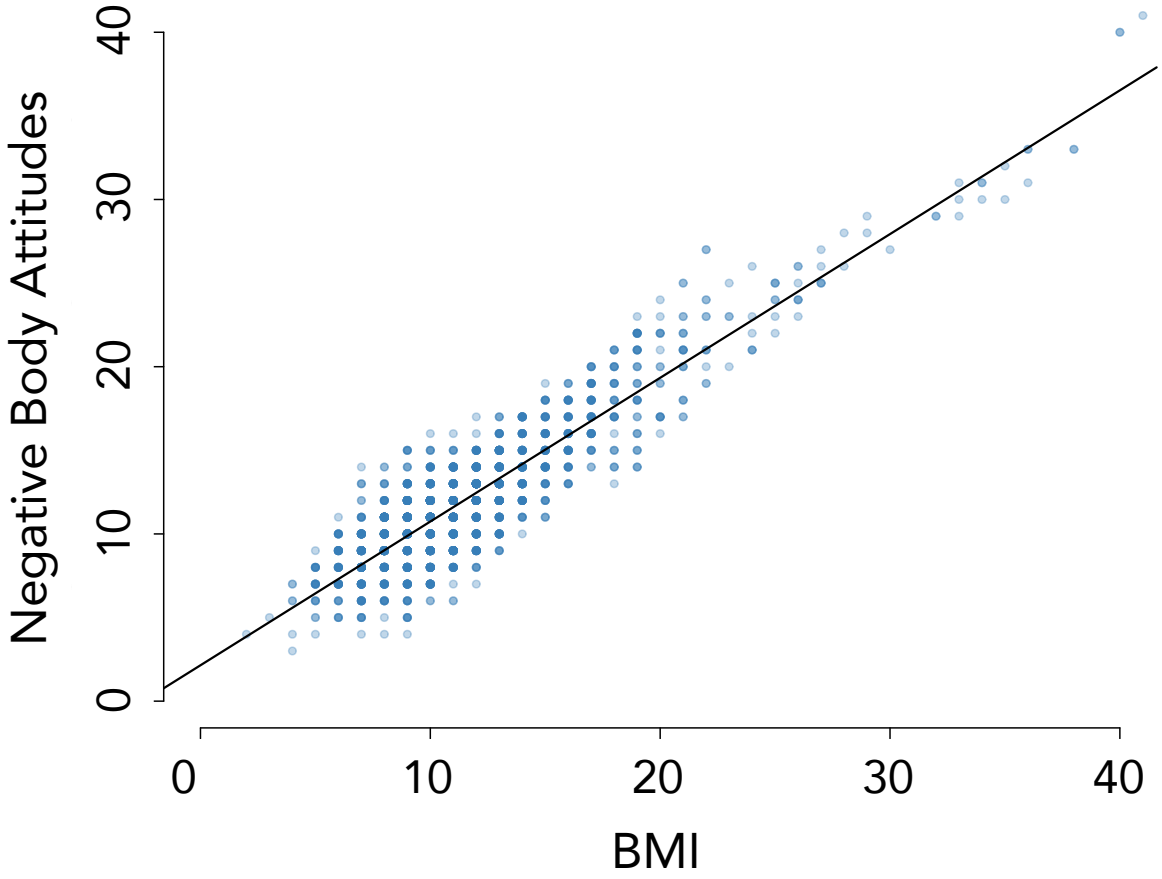
Iterative Recipe For Missing Data Estimation

1) Estimate parameters, treating imputations as real data

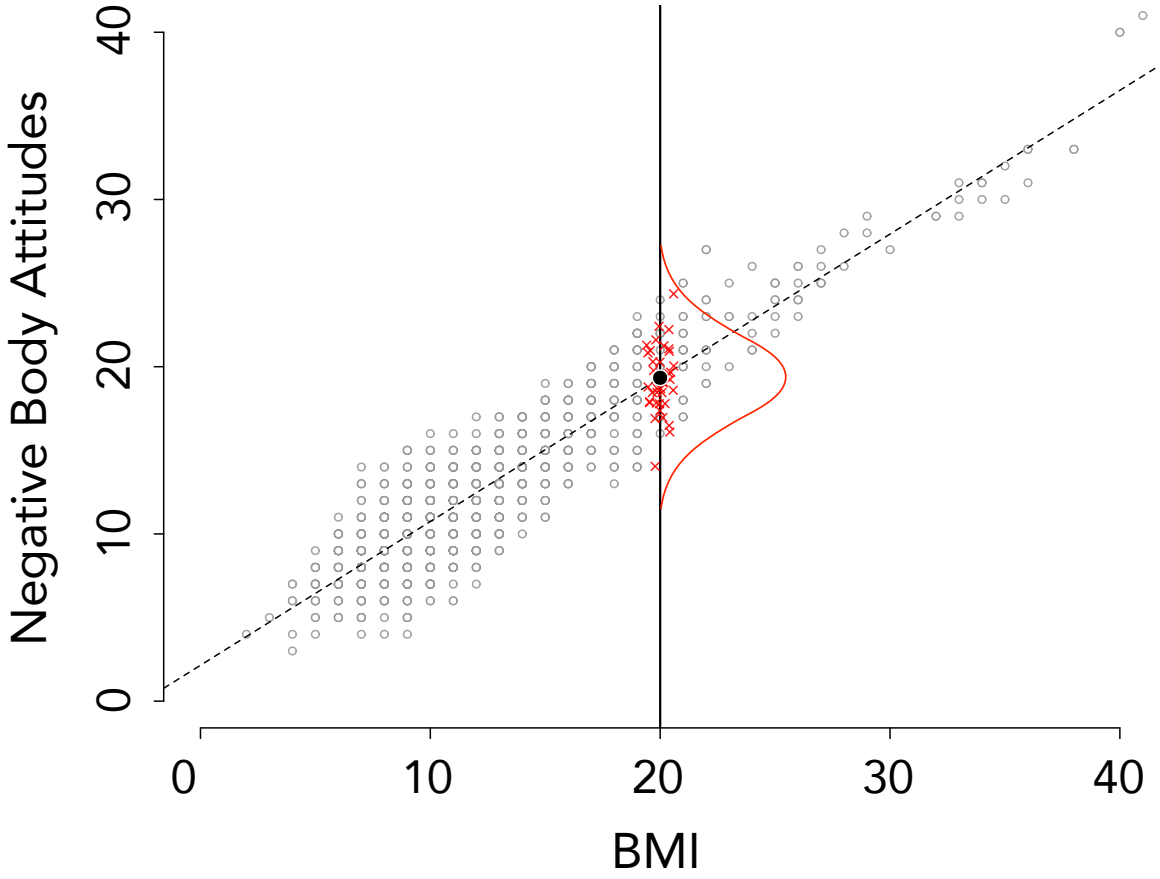
Model parameters  Missing data

2) Fill in missing values, treating parameters as true values

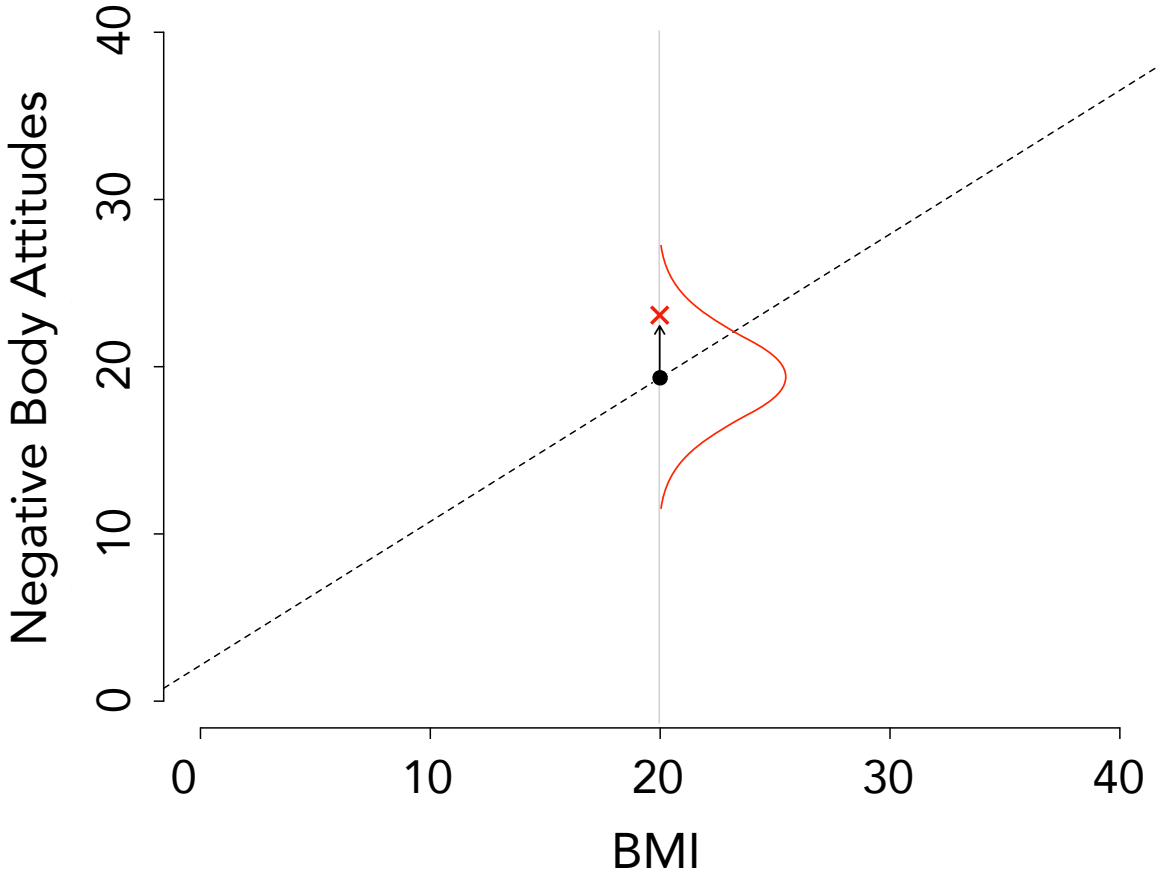
Bivariate Scatterplot



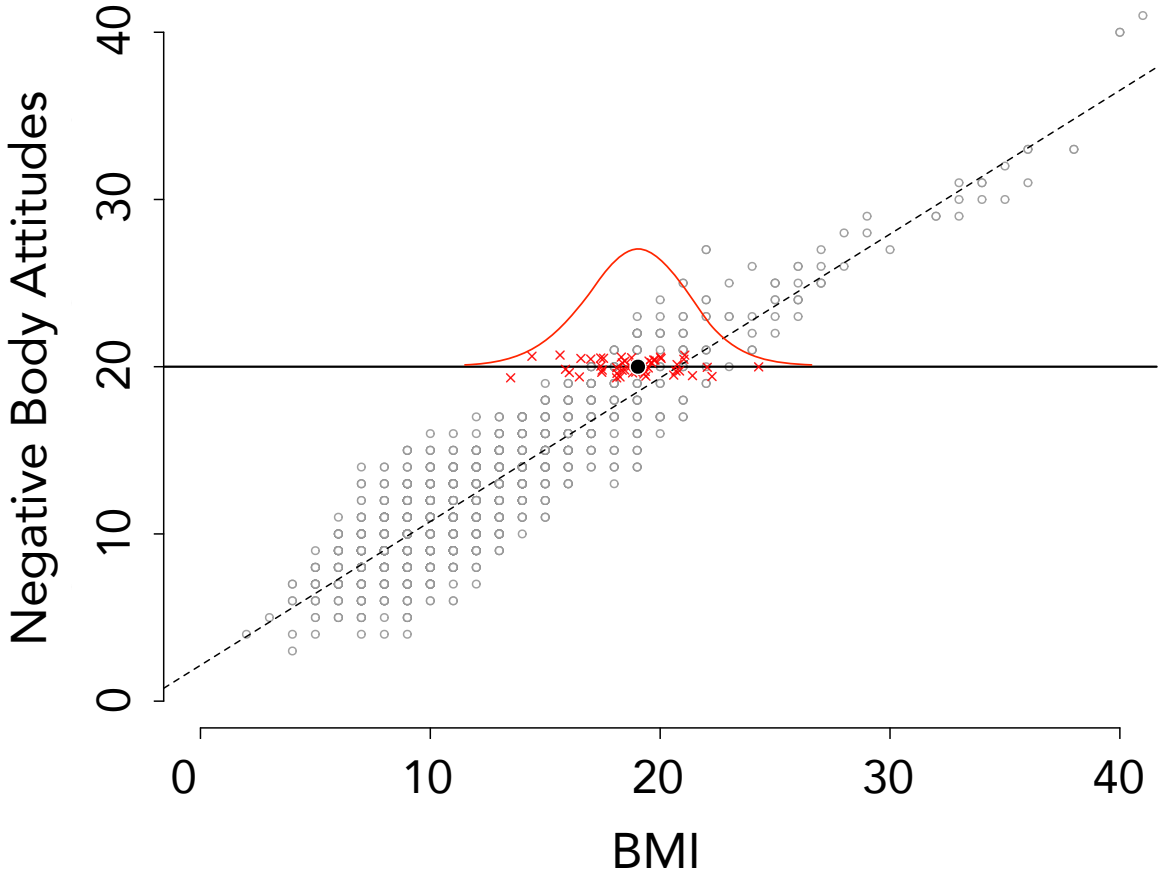
Distribution Of Attitudes Imputations



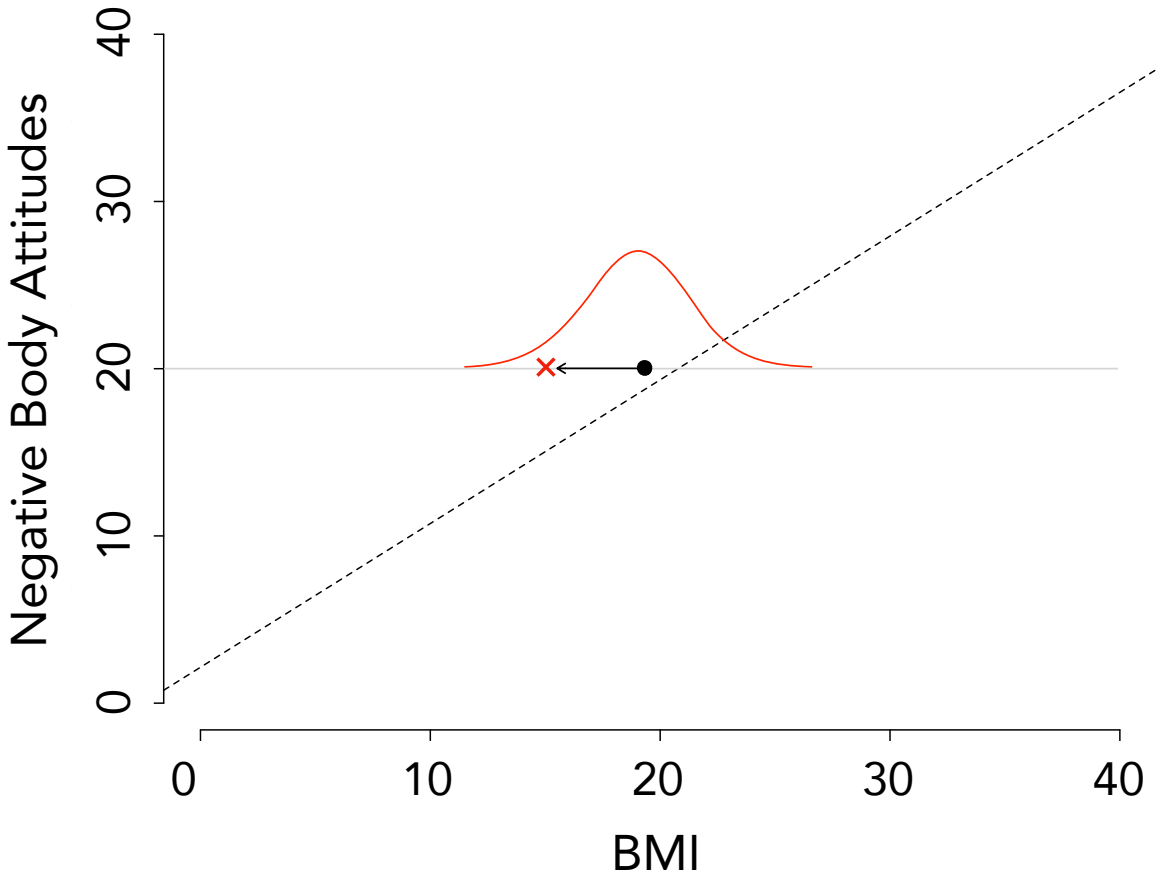
Imputation = Predicted Value + Noise



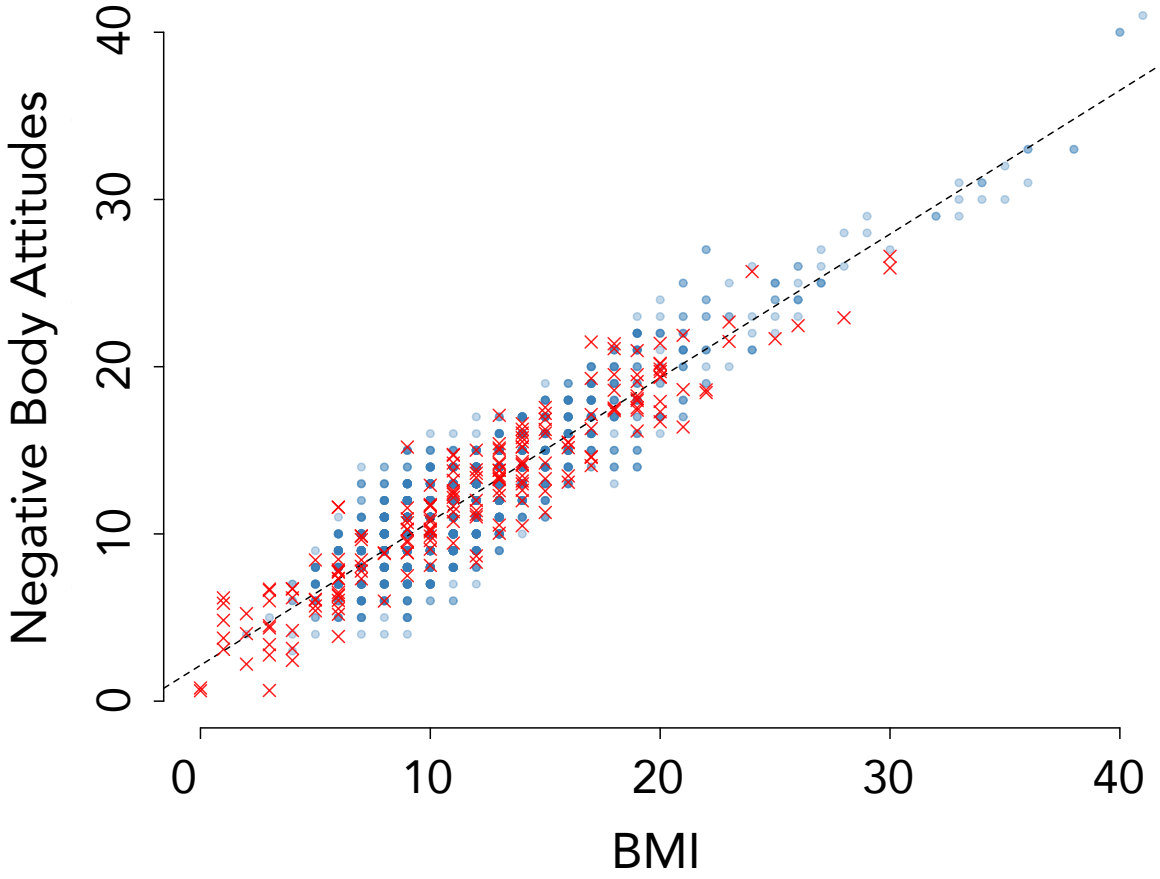
Distribution Of BMI Imputations



Imputation = Predicted Value + Noise



Imputed Data Set



Blimp Bayes Regression Script

```
DATA: bodyattitudes.dat;
VARIABLES: id bullied bmi batt1 batt2 batt3 batt4 batt5 negbodyatt;
NOMINAL: bullied;
MISSING: 999;
MODEL: negbodyatt ~ bmi bullied;
CENTER: bmi;
SEED: 90291;
BURN: 2000;
ITERATIONS: 10000;
OPTIONS: psr;
```

Categorical variable

Missing value code

Regression model

Center predictor

Random number seed (any integer)

MCMC iterations prior to analysis

MCMC iterations for analysis

MCMC diagnostic

Analysis Results

	Bayes		Maximum Likelihood		Multiple Imputation	
Parameter	**Est.**	**S.D.**	**Est.**	**S.E.**	**Est.**	**S.E.**
Intercept	19.780	0.209				
BMI	0.449	0.077				
BULLIED	4.487	0.638				

Mplus FIML Regression Script

```
DATA:
file = bodyattitudes.dat;
VARIABLE:
names = id bullied bmi batt1 batt2 batt3 batt4 batt5 negbodyatt;
usevariables = bullied bmi negbodyatt;
missing = all(999);
DEFINE:
center bmi (grandmean);
MODEL:
bmi bullied;
negbodyatt on bmi bullied;
OUTPUT:
stdyx;
```

← Select variables

← Missing value code

← Center predictor

← Predictor variables

← Regression model

← Standardized estimates

Analysis Results

	Bayes		Maximum Likelihood		Multiple Imputation	
Parameter	**Est.**	**S.D.**	**Est.**	**S.E.**	**Est.**	**S.E.**
Intercept	19.780	0.209	19.747	0.198		
BMI	0.449	0.077	0.457	0.076		
BULLIED	4.487	0.638	4.652	0.646		

Multiple Imputation Step 1: Create M Complete Data Sets

Y	X_1	X_2
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

Y	X_1	X_2
4	4	3
3	3.3	5
7	1	6
2.4	1	6
5	9	3
3	2.1	1.9
1	6	7
9	4	9
2	5.3	6

1

Y	X_1	X_2
4	4	3
3	4.7	5
7	1	6
1.3	1	6
5	9	3
3	6.5	3.5
1	6	7
9	4	9
2	4.2	6

2

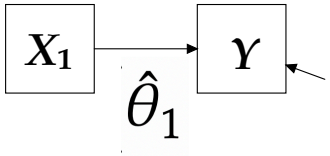
$\cdots$

Y	X_1	X_2
4	4	3
3	2.6	5
7	1	6
2.1	1	6
5	9	3
3	3.9	3.0
1	6	7
9	4	9
2	4.6	6

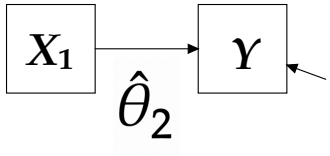
M

Multiple Imputation Step 2: Perform Analysis on Each Data Set

Y	X_1	X_2
4	4	3
3	3.3	5
7	1	6
2.4	1	6
5	9	3
3	2.1	1.9
1	6	7
9	4	9
2	5.3	6

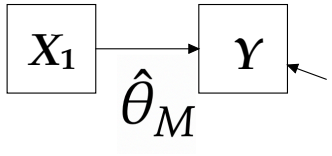


Y	X_1	X_2
4	4	3
3	4.7	5
7	1	6
1.3	1	6
5	9	3
3	6.5	3.5
1	6	7
9	4	9
2	4.2	6

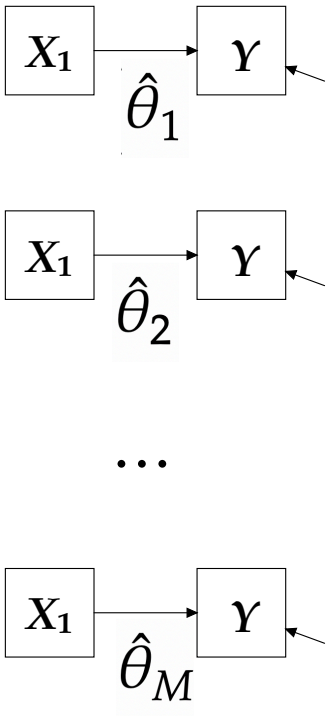


...

Y	X_1	X_2
4	4	3
3	2.6	5
7	1	6
2.1	1	6
5	9	3
3	3.9	3.0
1	6	7
9	4	9
2	4.6	6



Multiple Imputation Step 3: Combine Estimates and Standard Errors



$$\hat{\theta} = \left(\hat{\theta}_1 + \hat{\theta}_2 + \cdots + \hat{\theta}_M\right)/M$$

Blimp Multiple Imputation Script

```
DATA: bodyattitudes.dat;
VARIABLES: id bullied bmi batt1 batt2 batt3 batt4 batt5 negbodyatt;
NOMINAL: bullied;        <-- Categorical variable
MISSING: 999;            <-- Missing value code
FCS: negbodyatt bmi bullied;   <-- Variables to be imputed
SEED: 90291;             <-- Random number seed (any integer)
BURN: 2000;              <-- MCMC iterations prior to saving first data set
THIN: 2000;              <-- MCMC iterations between each additional data set
NIMPS: 20;               <-- Number of imputations
SAVE: separate = imputation*.dat;   <-- Stack imputed data sets into a file
OPTIONS: psr;            <-- MCMC diagnostic
```

Mplus Analyze Imputations Script

```
DATA:
file = imputationlist.dat;      <-- Text file with 20 data set names
type = imputation;              <-- Imputation data
VARIABLE:
names = id bullied bmi batt1 batt2 batt3 batt4 batt5 negbodyatt;
usevariables = bullied bmi negbodyatt;   <-- Select variables
DEFINE:
center bmi (grandmean);         <-- Center predictor
MODEL:
negbodyatt on bmi bullied;      <-- Regression model
OUTPUT:
stdyx;                          <-- Standardized estimates
```

Analysis Results

Methods are effectively equivalent!

	Bayes		Maximum Likelihood		Multiple Imputation	
Parameter	**Est.**	**S.D.**	**Est.**	**S.E.**	**Est.**	**S.E.**
Intercept	19.780	0.209	19.747	0.198	19.788	0.220
BMI	0.449	0.077	0.457	0.076	0.429	0.080
BULLIED	4.487	0.638	4.652	0.646	4.319	0.606


Q
A