

## Introduction to Missing Data Analyses

2020 Summer Institute  
for the UC Adolescence Consortium

Craig Enders  
University of California - Los Angeles  
cenders@psych.ucla.edu

1

## 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

2

## Negative Body Attitudes Questionnaire

	Strongly Disagree							Strongly Agree
1. My hips seem too 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	

3

## 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

4

## Substantive Analysis

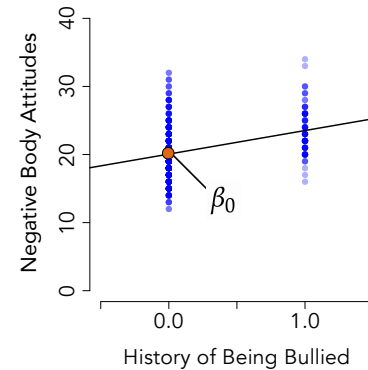
Negative body attitudes scale score regressed on body mass index and bullied indicator

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

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

5

## Regression Intercept

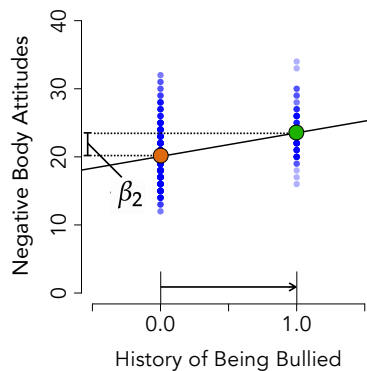


The intercept is the expected body attitudes score when bullied = 0

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

6

## 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

7

## 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

8

## Observed And Missing Data

Hypothetically complete data			Observed data			Missing data (latent, unseen)		
Y	X <sub>1</sub>	X <sub>2</sub>	Y	X <sub>1</sub>	X <sub>2</sub>	Y	X <sub>1</sub>	X <sub>2</sub>
4	4	3	4	4	3			
3	3	5	3	NA	5		3	
7	1	6	7	1	6			
2	1	6	NA	1	6	2		
5	9	3	5	9	3			
3	2	2	3	NA	NA		2	2
1	6	7	1	6	7			
9	4	9	9	4	9			
2	5	6	2	NA	6		5	

9

## 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

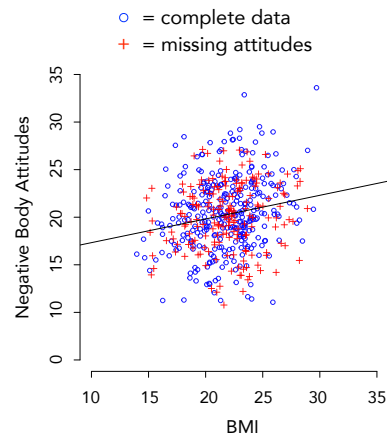
Y	X <sub>1</sub>	X <sub>2</sub>	Y	X <sub>1</sub>	X <sub>2</sub>
				NA	
			NA		
				NA	NA
					NA

10

## 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



11

## 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

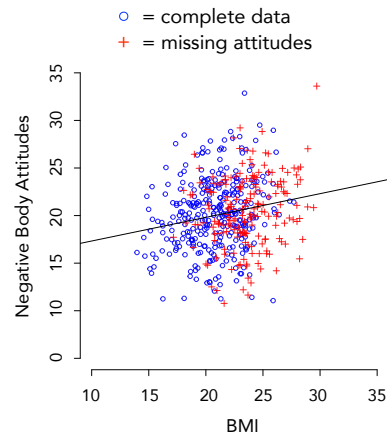
Y	X <sub>1</sub>	X <sub>2</sub>	Y	X <sub>1</sub>	X <sub>2</sub>
4	4	3			
3		5		NA	
7	1	6			
	1	6		NA	
5	9	3			
3				NA	NA
1	6	7			
9	4	9			
2		6		NA	

12

## 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

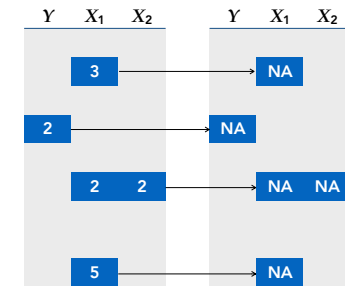


13

## 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

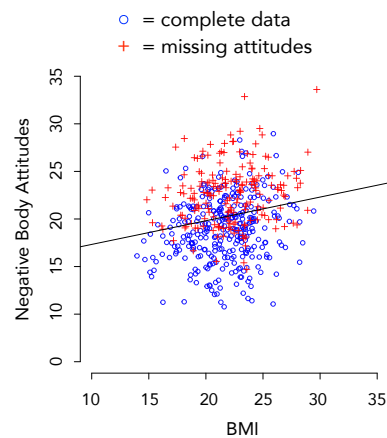


14

## 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



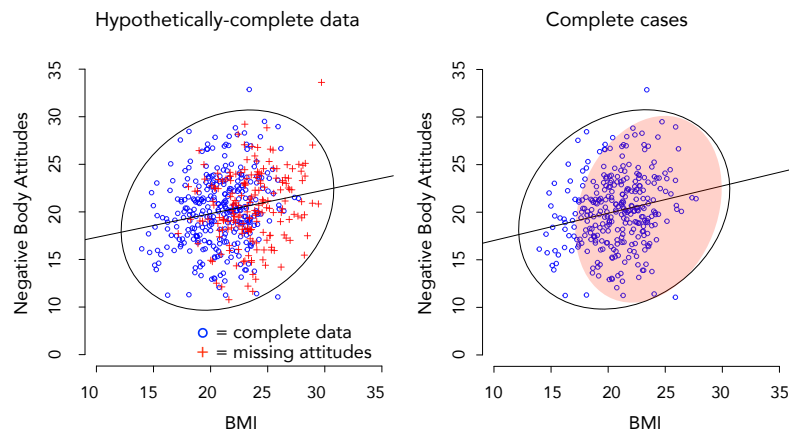
15

## 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

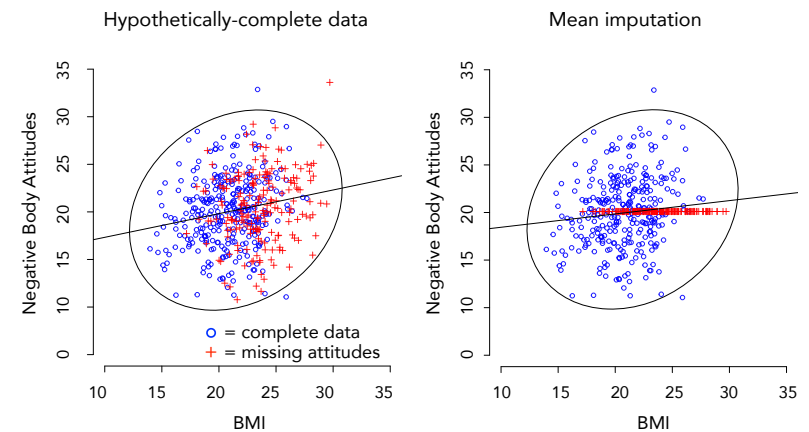
16

## Complete-Case Analysis (Deletion)



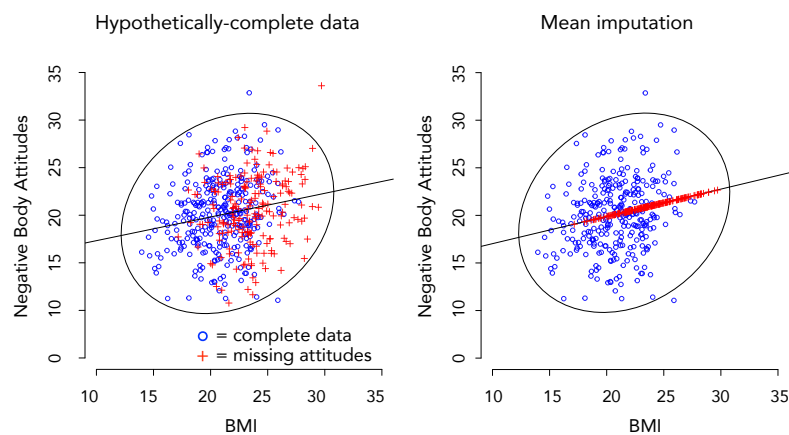
17

## Mean Imputation



18

## Regression Imputation



19

## Current Recommendations

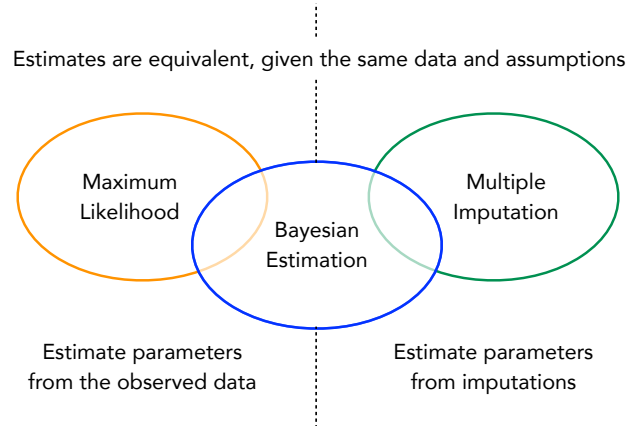
Methods that assume systematic non-response due to observed scores (MAR) are often optimal

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

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

20

## Relations Among Methods



21

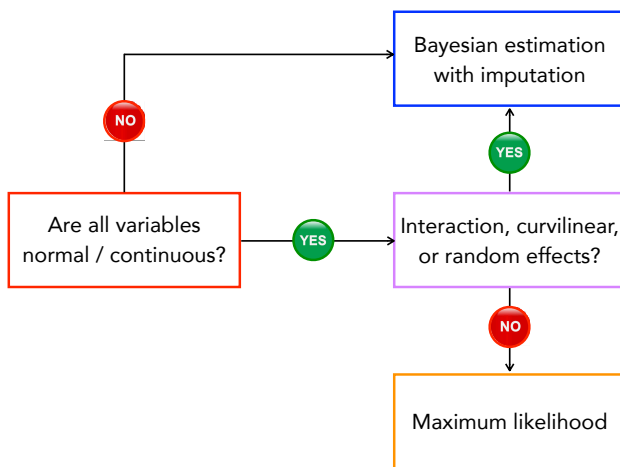
## Practical Benefits

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

22

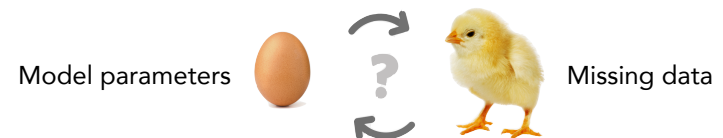
## Choosing A Missing Data Method



23

## Iterative Recipe For Missing Data Estimation

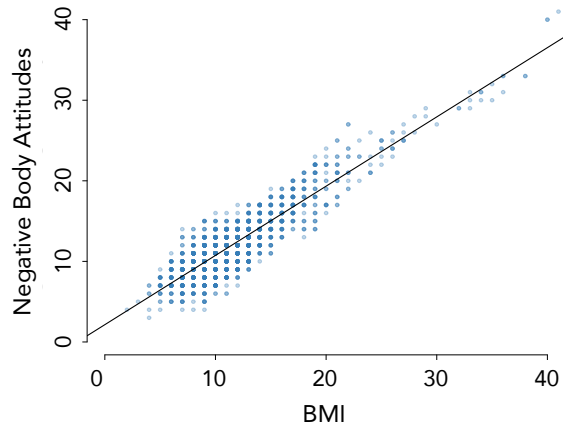
- 1) Estimate parameters, treating imputations as real data



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

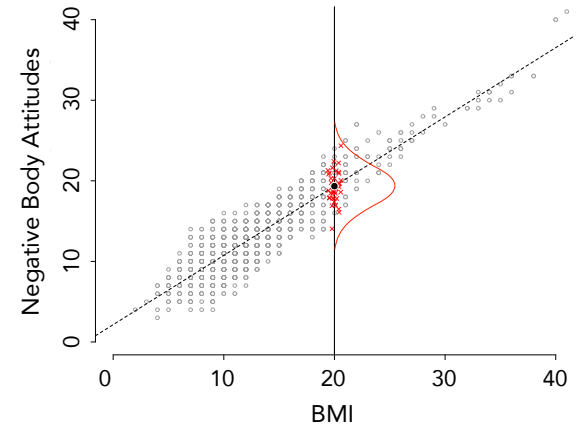
24

## Bivariate Scatterplot



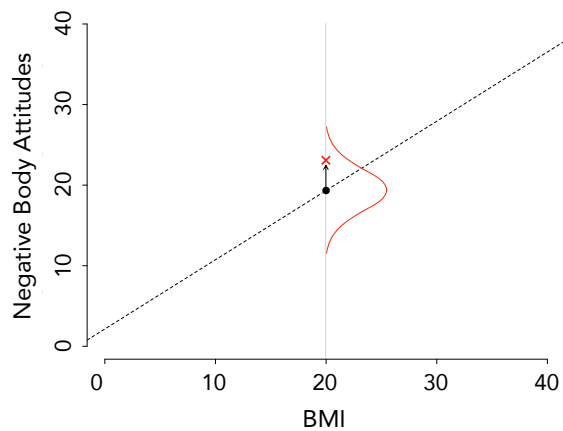
25

## Distribution Of Attitudes Imputations



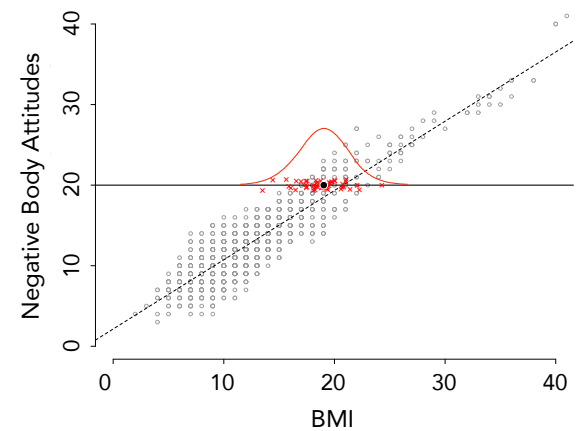
26

## Imputation = Predicted Value + Noise



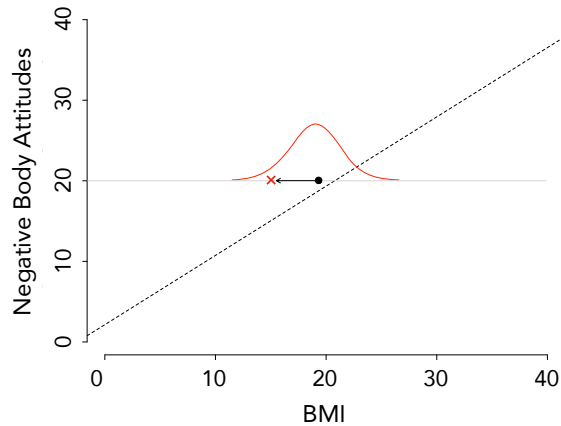
27

## Distribution Of BMI Imputations



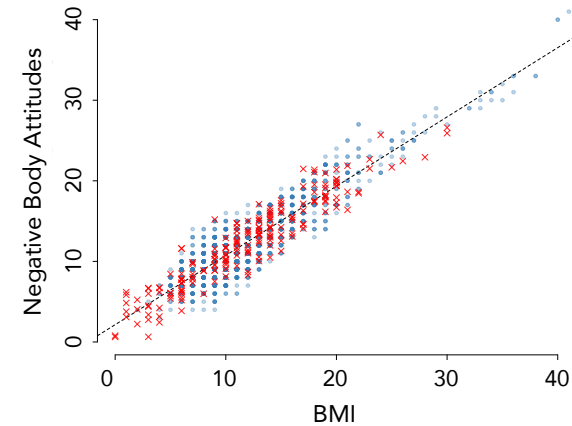
28

## Imputation = Predicted Value + Noise



29

## Imputed Data Set



30

## Blimp Bayes Regression Script

```
DATA: bodyattitudes.dat;
VARIABLES: id bullied bmi batt1 batt2 batt3 batt4 batt5 negbodyatt;
NOMINAL: bullied; ← Categorical variable
MISSING: 999; ← Missing value code
MODEL: negbodyatt ~ bmi bullied; ← Regression model
CENTER: bmi; ← Center predictor
SEED: 90291; ← Random number seed (any integer)
BURN: 2000; ← MCMC iterations prior to analysis
ITERATIONS: 10000; ← MCMC iterations for analysis
OPTIONS: psr; ← MCMC diagnostic
```

31

## Analysis Results

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

32



## Mplus FIML Regression Script

### DATA:

file = bodyattitudes.dat;

### VARIABLE:

names = id bullied bmi batt1 batt2 batt3 batt4 batt5 negbodyatt;

usevariables = bullied bmi negbodyatt; ← **Select variables**

missing = all(999); ← **Missing value code**

### DEFINE:

center bmi (grandmean); ← **Center predictor**

### MODEL:

bmi bullied; ← **Predictor variables**

negbodyatt on bmi bullied; ← **Regression model**

### OUTPUT:

stdyx; ← **Standardized estimates**

33

## Analysis Results

Parameter	Bayes		Maximum Likelihood		Multiple Imputation	
	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		

34

## Multiple Imputation Step 1: Create $M$ Complete Data Sets

Y	X <sub>1</sub>	X <sub>2</sub>
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 <sub>1</sub>	X <sub>2</sub>
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 <sub>1</sub>	X <sub>2</sub>
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

...

Y	X <sub>1</sub>	X <sub>2</sub>
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

35

## Multiple Imputation Step 2: Perform Analysis on Each Data Set

Y	X <sub>1</sub>	X <sub>2</sub>
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

$$X_1 \rightarrow \hat{\theta}_1 \rightarrow Y$$

Y	X <sub>1</sub>	X <sub>2</sub>
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

$$X_1 \rightarrow \hat{\theta}_2 \rightarrow Y$$

...

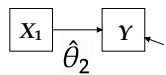
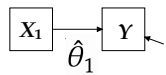
Y	X <sub>1</sub>	X <sub>2</sub>
4	4	3
3	2.6	5
7	1	6
2.1	1	6
5	9	3
3	3.9	3.0
1	6	7
9	4	9
2	4.6	6

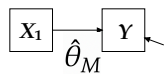
$$X_1 \rightarrow \hat{\theta}_M \rightarrow Y$$

36

## Multiple Imputation Step 3: Combine Estimates and Standard Errors



...



$$\hat{\theta} = (\hat{\theta}_1 + \hat{\theta}_2 + \dots + \hat{\theta}_M)/M$$

37

## 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

38

## 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

39

## Analysis Results

Methods are effectively equivalent!

Parameter	Bayes		Maximum Likelihood		Multiple Imputation	
	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

40

