Applied Missing Data Analysis

2017 ABCT Annual Convention
AMASS Workshop

Craig Enders
University of California - Los Angeles
Department of Psychology

Workshop Overview

Missing data mechanisms and assumptions

Multiple imputation and maximum likelihood estimation for normally distributed variables

Practical issues: categorical variables, composite variables, interaction effects, multilevel data

Analysis examples

Workshop Materials

Multiple imputation software available at www.appliedmissingdata.com/multilevel-imputation

Workshop slides and analysis scripts available at www.appliedmissingdata.com/training-materials

Analysis scripts for Mplus, R, SAS, SPSS, and Stata

Missing Data Mechanisms

Patterns Versus Mechanisms

The missing data pattern describes the configuration of observed and the missing values in a data set

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

The missing data mechanism describes how the probability of missingness is related to the data, if at all

Motivating Example

20 participants enroll in a smoking cessation study

Participants report age, number of years smoking, number of cigarettes smoked, and self-efficacy to quit

Number of cigarettes and self-efficacy have missing values

Age	Years	Cigs	Efficacy
29	7	9	NA
39	8	NA	NA
25	1	11	16
41	4	NA	21
39	6	10	17
41	8	5	10
46	8	7	13
40	10	NA	10
51	15	NA	11
43	5	11	13
26	9	12	11
51	11	11	16
36	14	NA	10
51	13	19	9
41	12	15	5
28	11	8	7
30	10	13	10
41	10	8	NA
23	7	10	7
33	11	10	6

Observed + Missing Data				Observed Data				Indicators	
Age	Years	Cigs	Efficacy	Age	Years	Cigs	Efficacy	Cigs	Efficacy
29	7	9	12	29	7	9	NA	0	1
39	8	12	14	39	8	NA	NA	1	1
25	1	11	16	25	1	11	16	0	0
41	4	3	21	41	4	NA	21	1	0
39	6	10	17	39	6	10	17	0	0
41	8	5	10	41	8	5	10	0	0
46	8	7	13	46	8	7	13	0	0
40	10	11	10	40	10	NA	10	1	0
51	15	12	11	51	15	NA	11	1	0
43	5	11	13	43	5	11	13	0	0
26	9	12	11	26	9	12	11	0	0
51	11	11	16	51	11	11	16	0	0
36	14	10	10	36	14	NA	10	1	0
51	13	19	9	51	13	19	9	0	0
41	12	15	5	41	12	15	5	0	0
28	11	8	7	28	11	8	7	0	0
30	10	13	10	30	10	13	10	0	0
41	10	8	15	41	10	8	NA	0	1
23	7	10	7	23	7	10	7	0	0
33	11	10	6	33	11	10	6	0	0

Notation and Terminology

The complete (hypothetical) data set is comprised of observed and missing parts, Y_{obs} and Y_{mis}

The unseen values in Y_{mis} can be viewed as latent scores

R is a missing data indicator (or matrix of indicators) where $R = 0$ if Y is observed and $R = 1$ if Y is missing

Missing Data Mechanisms

Rubin's (1976) missing data mechanisms describe relations between missing data indicators in R (the probability of nonresponse) and Y_{obs} and Y_{mis}

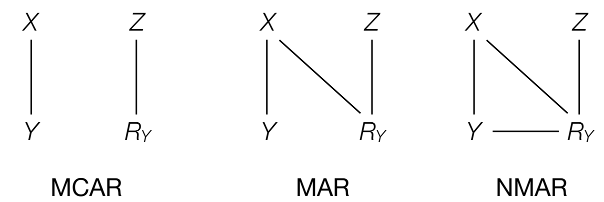
Missing completely at random (MCAR)

Missing at random (MAR)

Not missing at random (NMAR)

Diagram of Mechanisms

X represents a set of observed variables correlated with Y, Z represents a set of observed variables uncorrelated with X and Y, and R_Y is the missing data indicator for Y



Missing Completely At Random (MCAR)

The probability of missing data on a variable is unrelated to observed and latent parts of the data

$$P(R \mid Y_{obs}, Y_{mis}) = P(R)$$

MCAR implies an unsystematic process where all participants have the same chance of missing data

MCAR = observed and missing scores are the same, on average

Testing MCAR with Missing Data Indicators

MCAR is the only mechanism with testable propositions

Create a missing data indicator *R* for each incomplete variable (e.g., 0 = complete, 1 = missing) and examine mean differences across missing data patterns

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

MCAR Example

The absence of large mean differences supports MCAR

Pattern	Mean	SD	n
Age			
Complete	37.5	8.6	15
Missing	38.2	10.1	5
Years			
Complete	8.9	3.7	15
Missing	9.4	2.9	5
Self-Efficacy			
Complete	11.5	4.9	13
Missing	10.8	1.7	4

Age	Years	Cigs	SE	R_{Cigs}
29	7	9	NA	0
39	8	12	NA	0
25	1	11	16	0
41	4	3	21	0
39	6	10	17	0
41	8	5	10	0
46	8	7	13	0
40	10	11	10	0
51	15	12	11	0
43	5	NA	13	1
26	9	NA	11	1
51	11	11	16	0
36	14	10	10	0
51	13	NA	9	1
41	12	15	5	0
28	11	8	7	0
30	10	NA	10	1
41	10	NA	NA	1
23	7	10	7	0
33	11	10	6	0

Missing At Random (MAR)

The probability of missing data on a variable is unrelated to the latent parts of the data, but it can be related to the observed parts

$$P(R \mid Y_{obs}, Y_{mis}) = P(R \mid Y_{obs})$$

MAR implies systematic missingness where nonresponse varies across different observed score profiles

MAR = observed and missing scores are the same, on average, after conditioning on (controlling for) other variables

MAR Example

The presence of large mean differences refutes MCAR

Pattern	Mean	SD	n
Age			
Complete	36.5	9.4	15
Missing	41.4	5.7	5
Years			
Complete	8.1	3.2	15
Missing	11.8	2.9	5
Self-Efficacy			
Complete	12.0	4.5	13
Missing	9.0	2.7	4

Age	Years	Cigs	SE	R_{Cigs}
29	7	9	NA	0
39	8	NA	NA	1
25	1	11	16	0
41	4	3	21	0
39	6	10	17	0
41	8	5	10	0
46	8	7	13	0
40	10	NA	10	1
51	15	NA	11	1
43	5	11	13	0
26	9	12	11	0
51	11	11	16	0
36	14	NA	10	1
51	13	19	9	0
41	12	NA	5	1
28	11	8	7	0
30	10	13	10	0
41	10	8	NA	0
23	7	10	7	0
33	11	10	6	0

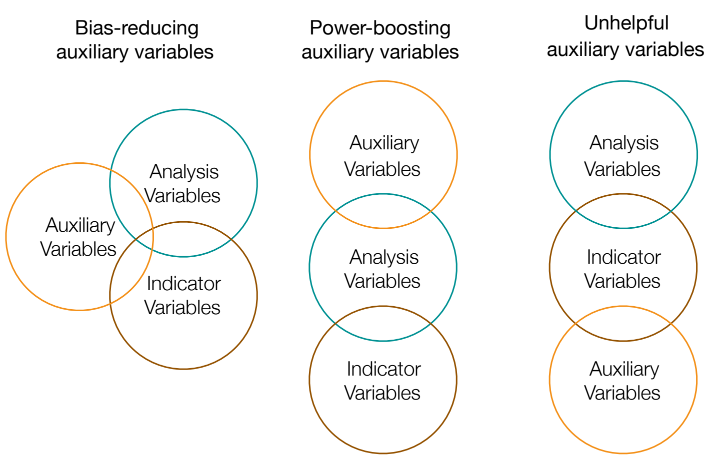
Inclusive Analysis Strategy and Auxiliary Variables

Satisfying MAR requires that we condition on all variables that simultaneously correlate with an incomplete variable and its missing data indicator

This may require additional auxiliary variables that wouldn't have appeared in the analysis had the data been complete

Choosing a small set of additional variables with the strongest correlations is a good strategy

Hierarchy of Auxiliary Variables



Not Missing At Random (NMAR)

The probability of missing data on a variable is related to the observed and latent parts of the data

$$P(R \mid Y_{obs}, Y_{mis})$$

NMAR implies systematic missingness where nonresponse depends on the latent (unseen) scores

NMAR = observed and missing scores are different, on average, after conditioning on (controlling for) other variables

NMAR Example

The presence of large mean differences refutes MCAR

Pattern	Mean	SD	n
Age			
Complete	35.5	8.5	15
Missing	44.4	6.1	5
Years			
Complete	8.1	3.3	15
Missing	11.6	2.7	5
Self-Efficacy			
Complete	12.1	4.5	13
Missing	8.8	2.6	4

Age	Years	Cigs	SE	R_{Cigs}
29	7	9	NA	0
39	8	NA	NA	1
25	1	11	16	0
41	4	3	21	0
39	6	10	17	0
41	8	5	10	0
46	8	7	13	0
40	10	NA	10	1
51	15	NA	11	1
43	5	11	13	0
26	9	12	11	0
51	11	11	16	0
36	14	10	10	0
51	13	NA	9	1
41	12	NA	5	1
28	11	8	7	0
30	10	13	10	0
41	10	8	NA	0
23	7	10	7	0
33	11	10	6	0

Why Mechanisms Matter

Mechanisms function as assumptions

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

Modern approaches like multiple imputation, maximum likelihood, and Bayes assume MAR (or MCAR)

Estimates are biased when assumptions are violated

Illustrative Computer Simulation

Generate 1000 samples of bivariate data with N = 250

Delete 50% of one variable’s scores according to an MCAR, MAR, or NMAR mechanism

Exclude incomplete cases or apply multiple imputation to each of the 1000 data sets

Compute the average estimates for both methods and compare to the true population values

MCAR Simulation Results

Both approaches are unbiased because assumptions about the nonresponse mechanism are satisfied

Parameter	True Value	Imputation	Deletion
Mean of X	100.00	100.02	99.98
Std. Dev. of X	13.00	12.97	13.38
Mean of Y	12.00	11.99	12.00
Std. Dev. of Y	3.00	2.99	3.00
Correlation	0.50	0.50	0.50

MAR Simulation Results

Deletion is biased because it assumes unsystematic nonresponse, imputation is accurate because the MAR assumption is satisfied

Parameter	True Value	Imputation	Deletion
Mean of X	100.00	100.01	110.35
Std. Dev. of X	13.00	12.98	7.86
Mean of Y	12.00	12.01	13.21
Std. Dev. of Y	3.00	2.99	2.76
Correlation	0.50	0.49	0.14

Bias Illustration

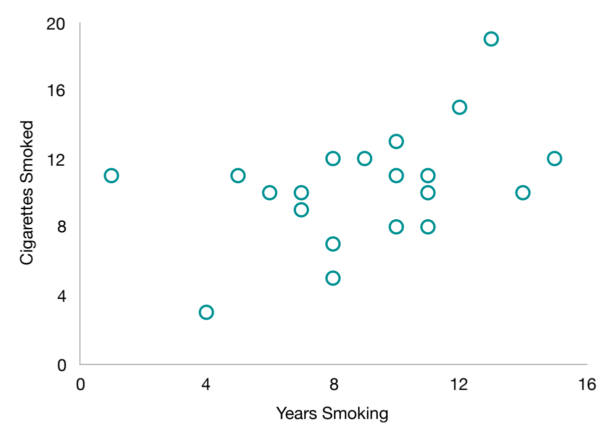
More years smoking is associated with higher rates of nonresponse (MAR)

Systematic missingness due to observed scores

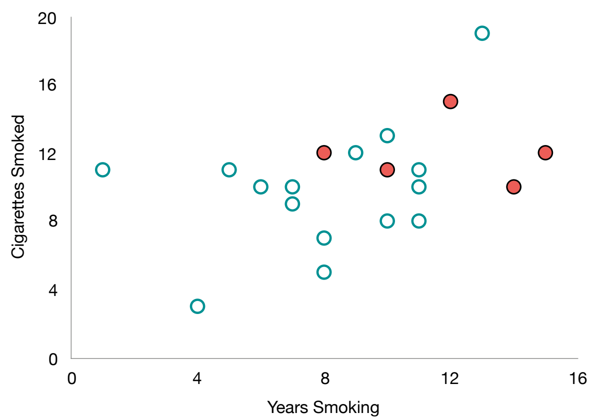
Deletion biases estimates because the complete cases are not representative of the full sample

Years	Cigs
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

Hypothetical Complete-Data Scatterplot

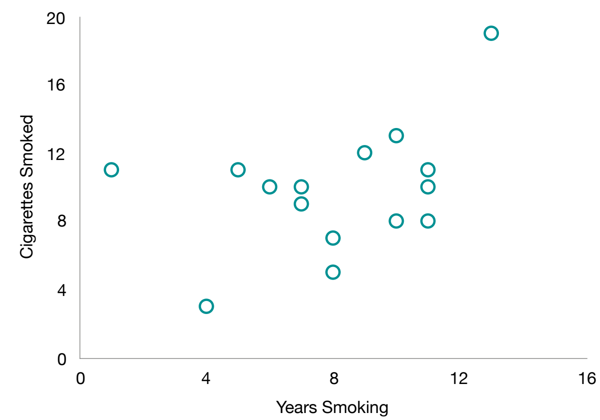


Deletion Scatterplot



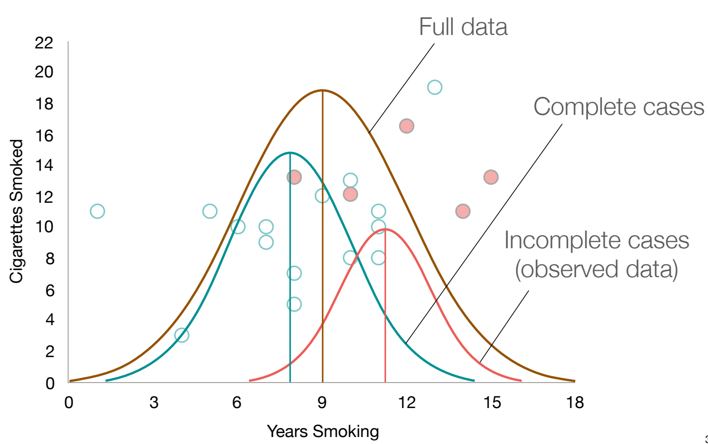
29

Deletion Scatterplot



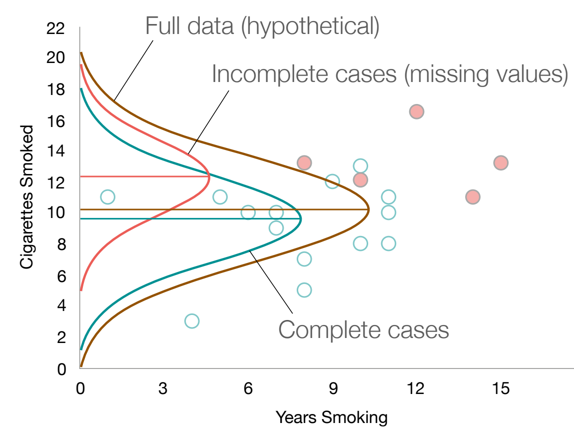
30

Distribution of Years Smoking



31

Distribution of Cigarettes Smoked



32

NMAR Simulation Results

Both methods are biased due to assumption violations, but deletion estimates are generally worse

Parameter	True Value	Imputation	Deletion
Mean of X	100.00	100.00	105.51
Std. Dev. of X	13.00	13.00	11.90
Mean of Y	12.00	14.12	14.40
Std. Dev. of Y	3.00	1.82	1.81
Correlation	0.50	0.36	0.32

Practical Recommendations

MAR-based methods are usually a good starting point but are not necessarily perfect solutions

We cannot test for an MAR mechanism and thus we must rely on logical arguments and knowledge about our data collection and participants to justify these methods

NMAR-based procedures are available but are difficult to implement and require other tenuous assumptions

Multiple Imputation

Multiple Imputation Overview

Multiple imputation generates several complete data sets (e.g., 20 or more), each with different imputations

Unique regression coefficients generate each data set

Analyzing multiple complete data sets provides a mechanism for adjusting standard errors

Multiple Imputation Steps: Imputation

The imputation phase creates multiple copies of the data, each with different replacement values

X	Y	Z
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

X	Y	Z
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	Y	Z
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	Y	Z
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 Steps: Analysis

In the analysis the researcher analyzes and obtains estimates from each complete data set

X	Y	Z
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

$X \xrightarrow{\theta_1} Y$

X	Y	Z
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

$X \xrightarrow{\theta_2} Y$

X	Y	Z
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

$X \xrightarrow{\theta_3} Y$

Multiple Imputation Steps: Pooling

The pooling phase combines the estimates and standard errors into a single set of results

$X \xrightarrow{\theta_1} Y$

$X \xrightarrow{\theta_2} Y$

$X \xrightarrow{\theta_3} Y$

$$\bar{\theta} = \frac{(\theta_1 + \theta_2 + \theta_3)}{3}$$

How Imputation Works

The imputation phase creates many imputed data sets, each generated from unique regression parameters

Imputation uses a regression model with an incomplete variable as the outcome and complete (and previously imputed) variables as predictors

Imputation = predicted score + random noise

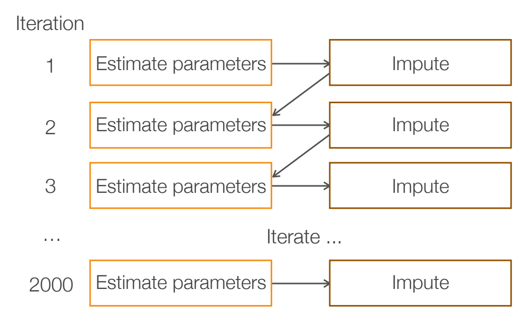
Markov Chain Monte Carlo (MCMC)

Imputation use an iterative MCMC algorithm that applies two major steps: estimate regression model parameters and update imputations based on the estimates

Bayesian estimation generates the regression parameters

The regression parameters define a distribution of plausible imputations for each observation

MCMC Algorithm for Imputation



Motivating Example

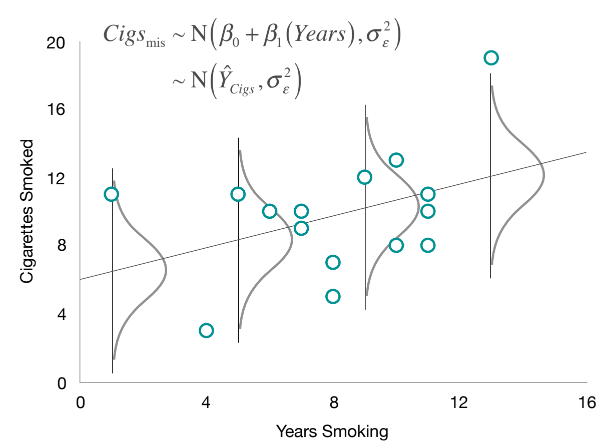
Number of years smoking and number of cigarettes smoked

More years smoking is associated with higher rates of nonresponse (MAR)

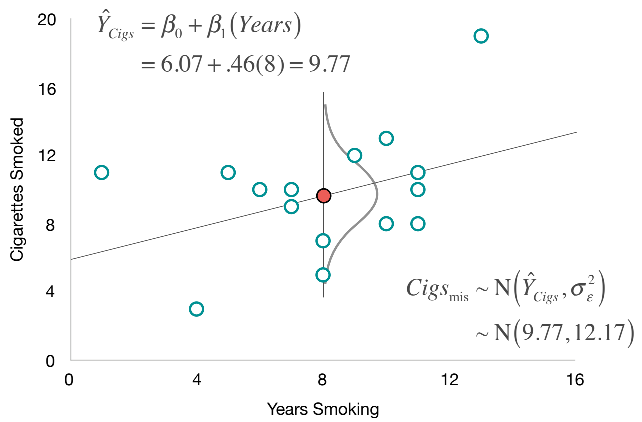
20% of respondents do not report the number of cigarettes smoked

Years	Cigs
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

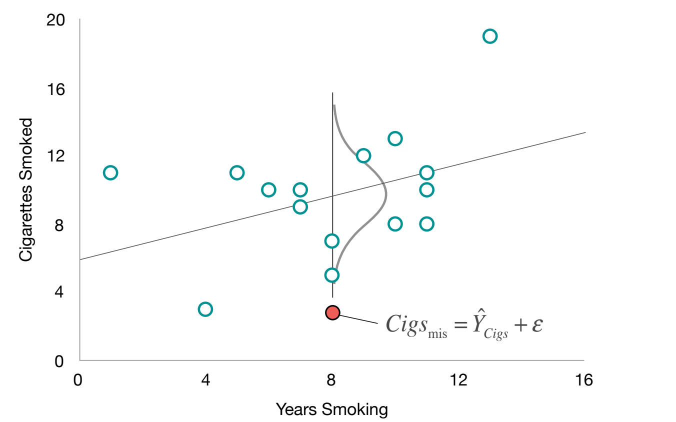
Distribution of Missing Values



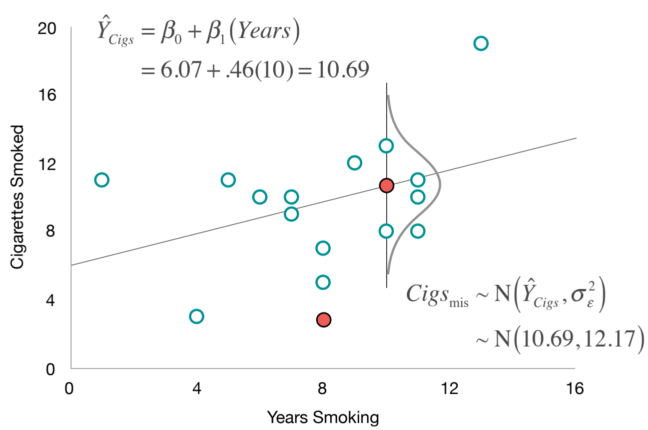
Imputation Step: Years = 8



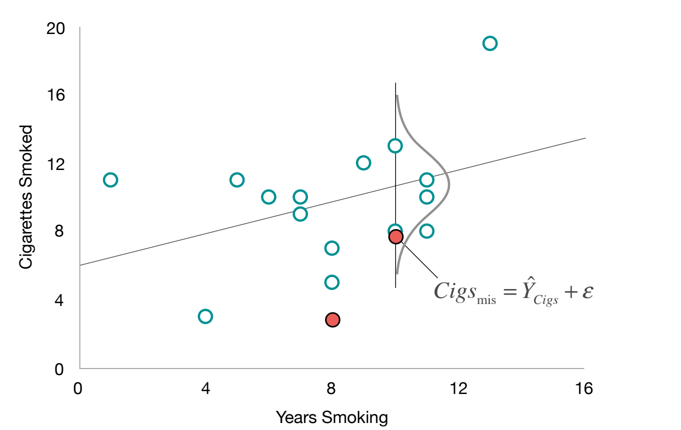
Imputation = Predicted Score + Noise



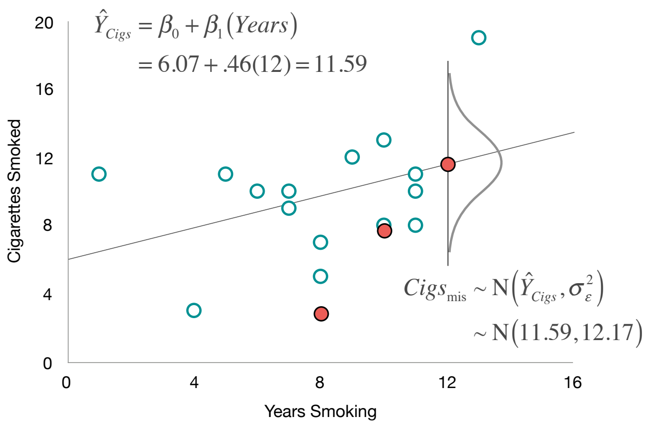
Imputation Step: Years = 10



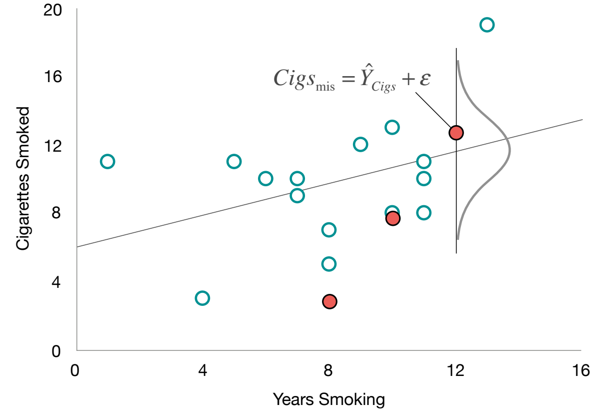
Imputation = Predicted Score + Noise



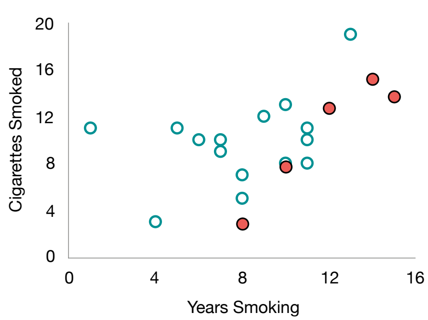
Imputation Example: Years = 12



Imputation = Predicted Score + Noise



Imputed Data



Years	Cigs
7	9
8	2.80
1	11
4	3
6	10
8	5
8	7
10	7.67
15	13.66
5	11
9	12
11	11
14	15.16
13	19
12	12.68
11	8
10	13
10	8
7	10
11	10

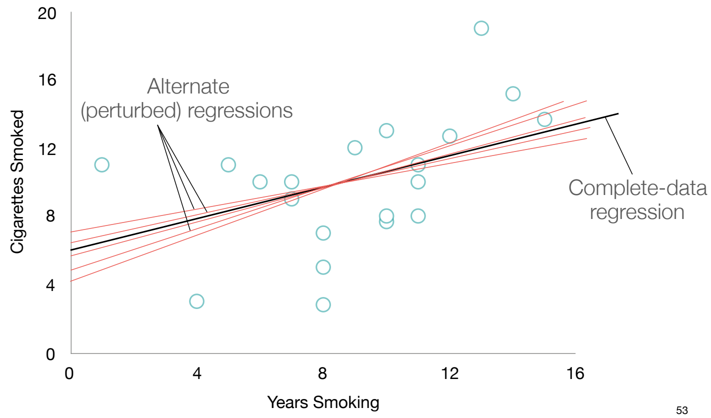
Updating Parameter Values

Bayesian estimation generates new regression parameters for the next round of imputation

Bayesian estimation "draws" new parameters from a posterior distribution (like a sampling distribution)

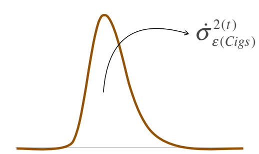
The updating process is akin to estimating the regression from the filled-in data and randomly perturbing estimates

Alternate Regression Lines



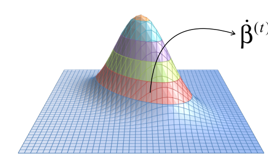
Bayesian Estimation Sequence

1. Draw new residual variance



$$\dot{\sigma}^{2(t)}_{\varepsilon} \sim \mathrm{IW}(df, SS_{\varepsilon})$$

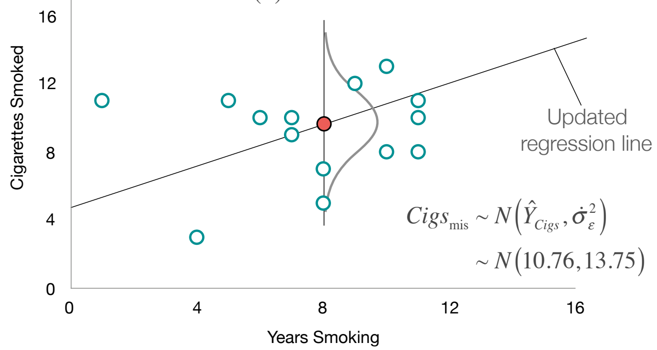
2. Draw new coefficients



$$\dot{\beta}^{(t)} \sim \mathrm{MVN}\left(\hat{\beta}, \Sigma_{\hat{\beta}}\right)$$

3. Update missing values with new parameters

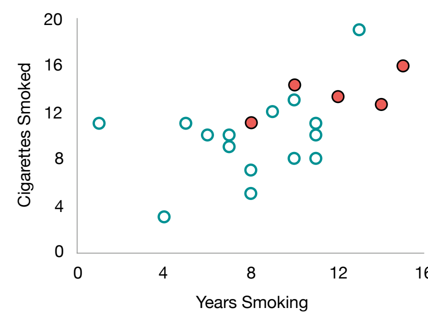
Next Imputation Step: Years = 8



$$\hat{Y}_{Cigs} = \dot{\beta}_0 + \dot{\beta}_1(Years) = 7.32 + .43(8) = 10.76$$

$$Cigs_{\text{mis}} \sim N\left(\hat{Y}_{Cigs}, \dot{\sigma}^2_{\varepsilon}\right) \sim N(10.76, 13.75)$$

Updated Imputations



Years	Cigs
7	9
8	11.05
1	11
4	3
6	10
8	5
8	7
10	14.28
15	15.91
5	11
9	12
11	11
14	12.61
13	19
12	13.28
11	8
10	13
10	8
7	10
11	10

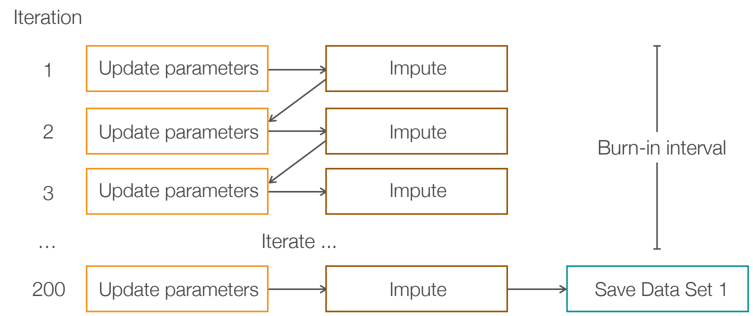
Burn-in and Thinning

The first imputed data set is saved only after a burn-in period where parameters achieve stable distributions

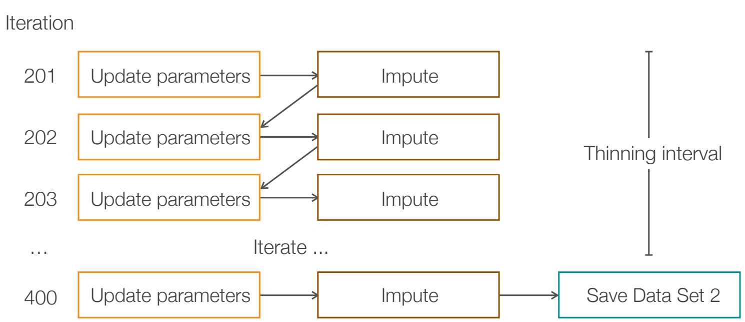
Imputed data sets from consecutive MCMC cycles are highly correlated (too similar) and so we want to allow MCMC cycles to lapse between each saved data set

Saving a data set at regular intervals (e.g., after every 200th imputation step) eliminates this autocorrelation

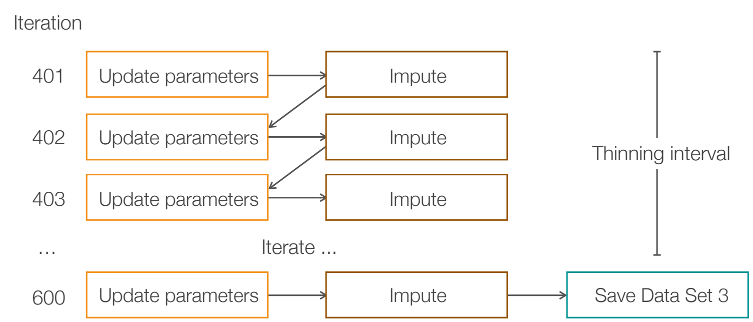
Burn-in Interval



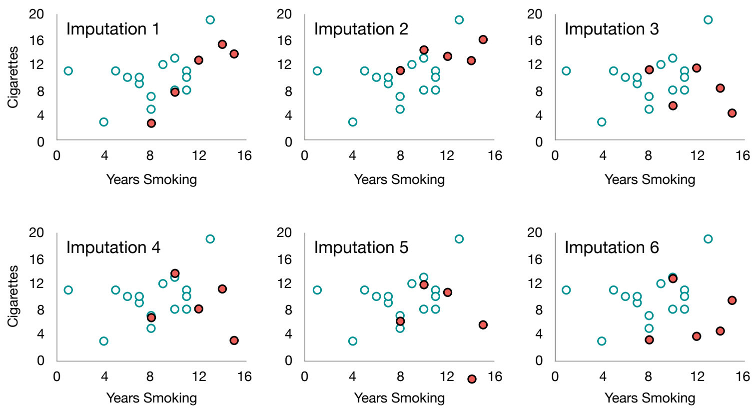
Thinning (Between-Imputation) Interval



Thinning Interval, Continued



Imputed Data Sets



Fully Conditional Specification (FCS) Imputation

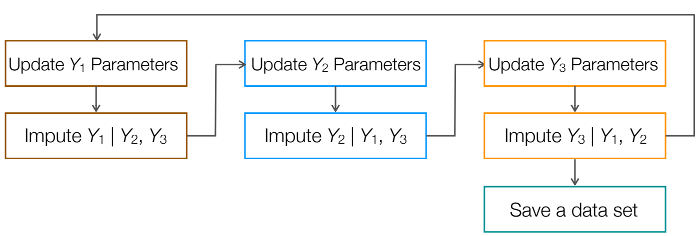
FCS imputes variables in a sequence, drawing missing values from a univariate normal distribution

FCS is just a series of univariate imputation problems

The incomplete variable from one step serves as a complete predictor in all other imputation steps

FCS Imputation Scheme

FCS imputation uses a series of univariate regression models to impute incomplete variables in a sequence



Blimp Script Ex0a.imp
Diagnostic Phase

```
DATA: ~/desktop/examples/smoking.dat;
VARNAMES: id txgroup txdum1 txdum2 male age years
  cigs heavycig efficacy stress;
MISSING: -99;
MODEL: ~ years cigs efficacy;
SEED: 90291;
BURN: 3000;
THIN: 1;
NIMPS: 2;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate psr;
CHAINS: 2 processors 2;
```

Potential Scale Reduction (PSR) Factors

The PSR captures the degree of similarity between imputations generated from two separate MCMC runs

The MCMC algorithm converges when the two runs begin to produce similar imputations

PSR < 1.05 to 1.10 is often considered acceptable

Use the PSR to specify the burn-in and thinning intervals

Diagnostic Output

```
POTENTIAL SCALE REDUCTION (PSR) OUTPUT:

 Comparing iterations 51 to 100 for 2 chains.
                     ---------------------------------------------------
                     |     Fix Eff| Ran Eff Var|     Err Var|   Threshold|
                     ---------------------------------------------------
          Max PSR    |       1.027|         nan|       1.008|         nan|
 Missing Variable    |        cigs|            |        cigs|            |
                     ---------------------------------------------------

 Comparing iterations 101 to 200 for 2 chains.
                     ---------------------------------------------------
                     |     Fix Eff| Ran Eff Var|     Err Var|   Threshold|
                     ---------------------------------------------------
          Max PSR    |       1.043|         nan|       1.002|         nan|
 Missing Variable    |    efficacy|            |        cigs|            |
                     ---------------------------------------------------
```


Blimp Script Ex0b.imp Imputation Phase (Mplus Format)

```
DATA: ~/desktop/examples/smoking.dat;
VARNAMES: id txgroup txdum1 txdum2 male age years
  cigs heavycig efficacy stress;
MISSING: -99;
MODEL: ~ years cigs efficacy;
SEED: 90291;
BURN: 100;
THIN: 100;
NIMPS: 20;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate;
CHAINS: 2 processors 2;
```

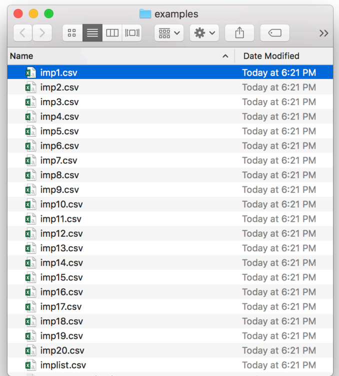

Blimp Script Ex0c.imp Imputation Phase (R, SAS, SPSS, and Stata Format)

```
DATA: ~/desktop/examples/smoking.dat;
VARNAMES: id txgroup txdum1 txdum2 male age years
  cigs heavycig efficacy stress;
MISSING: -99;
MODEL: ~ years cigs efficacy;
SEED: 90291;
BURN: 100;
THIN: 100;
NIMPS: 20;
OUTFILE: ~/desktop/examples/imps.csv;
OPTIONS: stacked;
CHAINS: 2 processors 2;
```

Imputed Data Sets

The imputation phase generates a set of imputed data sets

The next step is to analyze the data …



Multiple Imputation Analysis and Pooling Phase

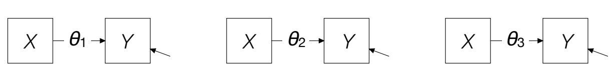
Multiple Imputation Steps: Analysis

In the analysis the researcher analyzes and obtains estimates from each complete data set

X	Y	Z
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

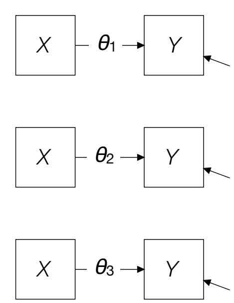
X	Y	Z
4	4	3
3	4.7	5
7	1	6
1.3	1	6
5	9	3
3	6.5	3.5
1	6	7
9	4	9
2	4.2	6

X	Y	Z
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 Steps: Pooling

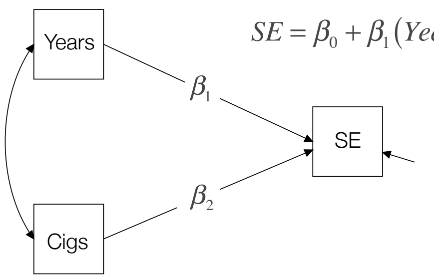
The pooling phase combines the estimates and standard errors into a single set of results



$$\bar{\theta} = \frac{(\theta_1 + \theta_2 + \theta_3)}{3}$$

Analysis Model

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



$$SE = \beta_0 + \beta_1(Years) + \beta_2(Cigs) + \varepsilon$$

Years	Cigs	Efficacy
7	9	15.50
8	8.96	15.78
1	11	16
4	3	21
6	10	17
8	5	10
8	7	13
10	9.92	10
15	13.62	11
5	11	13
9	12	11
11	11	16
14	14.42	10
13	19	9
12	18.04	5
11	8	7
10	13	10
10	8	9.18
7	10	7
11	10	6

Years	Cigs	Efficacy
7	9	15.38
8	8.28	9.25
1	11	16
4	3	21
6	10	17
8	5	10
8	7	13
10	13.41	10
15	6.99	11
5	11	13
9	12	11
11	11	16
14	15.31	10
13	19	9
12	12.75	5
11	8	7
10	13	10
10	8	14.85
7	10	7
11	10	6

Years	Cigs	Efficacy
7	9	5.86
8	10.04	13.88
1	11	16
4	3	21
6	10	17
8	5	10
8	7	13
10	12.95	10
15	14.40	11
5	11	13
9	12	11
11	11	16
14	14.47	10
13	19	9
12	11.46	5
11	8	7
10	13	10
10	8	6.79
7	10	7
11	10	6

Pooling Estimates

The multiple imputation point estimate is the arithmetic average of the *M* complete-data estimates

Estimate from one data set

Pooled estimate

$$\hat{\theta} = \frac{\sum_{m=1}^{M} \hat{\theta}_m}{M}$$

Number of data sets

Example: Descriptives

Data Set 1

	M	SD	N
Years	9.00	3.45	20
Cigs	10.59	3.84	20
SE	11.62	4.19	20

Data Set 2

	M	SD	N
Years	9.00	3.45	20
Cigs	10.19	3.61	20
SE	11.57	4.14	20

Data Set 3

	Years	Cigs	SE
Years	9.00	3.45	20
Cigs	10.52	3.51	20
SE	10.93	4.28	20

Pooled Estimates

	M	SD	N
Years	9.00	3.45	20
Cigs	10.43	3.65	20
SE	11.37	4.20	20

$$\hat{\theta} = \frac{\sum_{m=1}^{M} \hat{\theta}_m}{M} = \frac{10.59 + 10.19 + 10.52}{3} = 10.43$$

Example: Correlations

Data Set 1

	Years	Cigs	SE
Years	1.00		
Cigs	0.54	1.00	
SE	-0.60	-0.45	1.00

Data Set 2

	Years	Cigs	SE
Years	1.00		
Cigs	0.38	1.00	
SE	-0.57	-0.38	1.00

Data Set 3

	Years	Cigs	SE
Years	1.00		
Cigs	0.54	1.00	
SE	-0.52	-0.26	1.00

$$\hat{\theta} = \frac{\sum_{m=1}^{M} \hat{\theta}_m}{M} = \frac{-.45 - .57 - .26}{3} = -.37$$

Pooled Estimates

	Years	Cigs	SE
Years	1.00		
Cigs	0.49	1.00	
SE	-0.57	-0.37	1.00

Example: Regression Parameters

	Imp 1	Imp 2	Imp 3	Pooled
B_0 (Intercept)	19.20	19.16	18.30	18.88
B_1 (Years)	-0.62	-0.59	-0.63	-0.61
B_2 (Cigarettes)	-0.19	-0.22	0.04	-0.12
Residual Variance	12.07	12.41	14.81	13.10
R^2	0.39	0.35	0.28	0.34

$$\hat{\theta} = \frac{\sum_{m=1}^{M} \hat{\theta}_m}{M} = \frac{-.19 - .22 + .04}{3} = -.12$$

Pooling Standard Errors

Averaging standard errors underestimates sampling variability because the component standard errors are based on complete data sets

Imputation standard errors consist of two components

Within-imputation variance estimates complete-data sampling error, and between-imputation variance captures additional noise from the missing data

Within-Imputation Variance

The within-imputation variance is the average squared standard error

$$\mathrm{V_W} = \frac{\sum_{m=1}^{M} SE_m^2}{M}$$

Within-imputation variance estimates sampling error in the hypothetically complete data

Example

Data Set 1			
	Est.	**SE**	**SE²**
B_0	19.20	2.559	6.548
B_1	-0.62	0.275	0.076
B_2	-0.19	0.247	0.061

Data Set 2			
	Est.	**SE**	**SE²**
B_0	19.16	2.762	7.629
B_1	-0.59	0.253	0.064
B_2	-0.22	0.242	0.059

Data Set 3			
	Est.	**SE**	**SE²**
B_0	16.54	2.970	8.821
B_1	-0.67	0.304	0.092
B_2	0.04	0.299	0.089

$$V_W = \frac{\sum_{m=1}^{M} SE_m^2}{M} = \frac{.061 + .059 + .089}{3} = .097$$

Pooled Estimates		
	Est.	**V_W**
B_0	18.30	10.745
B_1	-0.63	0.080
B_2	-0.12	0.097

Between-Imputation Variance

Variability in the estimates across data sets results from using different imputations

$$V_B = \frac{\sum_{m=1}^{M} \left(\theta_m - \bar{\theta}\right)^2}{M - 1}$$

Between-imputation variance captures this additional variability by applying the sample variance formula to the *M* estimates

Example

Data Set 1			
	Est.	**SE**	**SE²**
B_0	19.20	2.559	6.548
B_1	-0.62	0.275	0.076
B_2	-0.19	0.247	0.061

Data Set 2			
	Est.	**SE**	**SE²**
B_0	19.16	2.762	7.629
B_1	-0.59	0.253	0.064
B_2	-0.22	0.242	0.059

Data Set 3			
	Est.	**SE**	**SE²**
B_0	16.54	2.970	8.821
B_1	-0.67	0.304	0.092
B_2	0.04	0.299	0.089

$$V_B = \frac{(-.19 + .12)^2 + (-.22 + .12)^2 + (.04 + .12)^2}{2}$$
$$= .021$$

Pooled Estimates			
	Est.	**V_W**	**V_B**
B_0	18.30	10.745	2.311
B_1	-0.63	0.080	0.002
B_2	-0.12	0.097	0.021

Total Variance and Standard Error

The total variance (squared standard error) combines complete-data sampling error and missing data uncertainty

Complete data sampling error

Missing data uncertainty

Sampling variance of the mean estimate

$$V_T = V_W + V_B + \frac{V_B}{M}$$

$$SE = \sqrt{V_W + V_B + \frac{V_B}{M}} = \sqrt{V_T}$$

Significance Testing

The usual t or z ratio is based on pooled quantities

Pooled estimate

Hypothesized value

$$t\,(\text{or } z) = \frac{\hat{\theta} - \theta_0}{SE}$$

Pooled standard error

Multivariate significance tests (e.g., Wald and likelihood ratio) are also available

Mplus Script Ex0a.inp Analysis and Pooling Phase

```
DATA:
file = implist.csv;
type = imputation;
VARIABLE:
names = id txgroup txdum1 txdum2 male age years
  cigs heavycig efficacy stress;
usevariables = years cigs efficacy;
MODEL:
efficacy on years (b1)
  cigs (b2);
MODEL TEST:
b1 = 0; b2 = 0;
OUTPUT:
standardized(stdyx);
```


Mplus Output

```
MODEL RESULTS

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value
 EFFICACY ON
    YEARS             -0.652      0.269     -2.423      0.015
    CIGS              -0.069      0.278     -0.249      0.803
 Intercepts
    EFFICACY          17.815      2.818      6.323      0.000
 Residual Variances
    EFFICACY          11.240      3.972      2.830      0.005
```


Mplus Output, Continued

```
STANDARDIZED MODEL RESULTS

STDYX Standardization

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value
 EFFICACY ON
    YEARS             -0.540      0.201     -2.684      0.007
    CIGS              -0.055      0.254     -0.216      0.829
 Intercepts
    EFFICACY           4.393      0.735      5.975      0.000
 Residual Variances
    EFFICACY           0.677      0.178      3.791      0.000

R-SQUARE
    Observed                                        Two-Tailed
    Variable        Estimate       S.E.  Est./S.E.    P-Value
    EFFICACY           0.323      0.178      1.813      0.070
```

Maximum Likelihood Estimation

Maximum Likelihood Overview

Maximum likelihood identifies the population parameter values that best fit the observed data

The analysis uses the incomplete data, and missing data handling is integrated into estimation

An iterative algorithm generates temporary imputations at each computational cycle as it searches for the optimal parameter values (i.e., implicit imputation)

Multivariate Normal Distribution

Multivariate normal distribution function

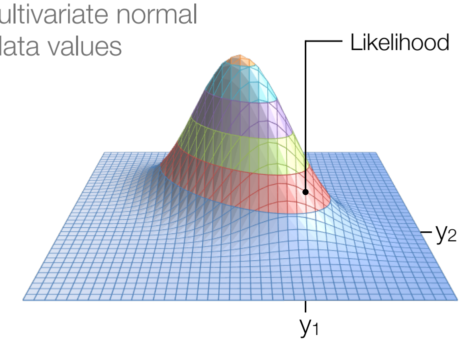
$\{\ldots\}$ = Likelihood

$$\log L_i = \log\left\{\frac{1}{\sqrt{2\pi}^{k/2}\left|\Sigma\right|^{.5}} e\left[-.5\left(\mathbf{y}_i - \mu\right)^T \Sigma^{-1}\left(\mathbf{y}_i - \mu\right)\right]\right\}$$

The likelihood gives the probability that a set of scores came from a multivariate normal distribution with a particular mean vector and covariance matrix

Geometric Interpretation

The likelihood expression returns the height of the multivariate normal distribution at the data values



Mahalanobis Distance

The key kernel of the likelihood is a squared z-score that gives the sum of squared standardized deviation scores

Deviation scores for observation *i*

$$z_i^2 = \left(\mathbf{y}_i - \mu\right)^T \Sigma^{-1} \left(\mathbf{y}_i - \mu\right)$$

Standardize by "dividing by" the covariance matrix

Sample Log Likelihood

The log likelihood gives the probability of the sample data, given a multivariate normal distribution with a particular mean vector and covariance matrix

$$\log L = \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi}^{k/2} \left|\Sigma\right|^{.5}} e\left[-.5\left(\mathbf{y}_i - \mu\right)^T \Sigma^{-1} \left(\mathbf{y}_i - \mu\right) \right] \right)$$

$$= \sum_{i=1}^{N} \left(-\frac{k}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left|\Sigma\right| - \frac{1}{2}\left(\mathbf{y}_i - \mu\right)^T \Sigma^{-1} \left(\mathbf{y}_i - \mu\right) \right)$$

Missing-Data Log Likelihood

Number of observed scores for case *i*

Mean vector elements corresponding to observed scores

$$\log L = \sum_{i=1}^{N} \left(-\frac{k_i}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left|\Sigma_i\right| - \frac{1}{2}\left(\mathbf{y}_i - \mu_i\right)^T \Sigma_i^{-1} \left(\mathbf{y}_i - \mu_i\right) \right)$$

Observed scores for case *i*

Covariance matrix elements corresponding to observed scores

Missing-Data Log Likelihood, Continued

The squared z-scores in the log likelihood are compute using all available data

Deviation scores are computed using only the means that correspond to the observed data for an observation

The size of the covariance matrix used to standardize the deviation scores adjusts to the observed data

Motivating Example

Number of years smoking and number of cigarettes smoked

More years smoking is associated with higher rates of nonresponse (MAR)

Two missing data patterns

Years (Y_1)	Cigs (Y_2)
1	11
4	3
5	11
6	10
7	9
7	10
8	5
8	7
9	12
10	8
10	13
11	8
11	10
11	11
13	19
8	NA
10	NA
12	NA
14	NA
15	NA

Mahalanobis Distance Computations

Complete cases

$$z_i^2 = (\mathbf{y}_i - \mu_i)^T \Sigma_i^{-1} (\mathbf{y}_i - \mu_i)$$

$$= \left(\begin{bmatrix} Y_{1i} \\ Y_{2i} \end{bmatrix} - \begin{bmatrix} \mu_{Y_1} \\ \mu_{Y_2} \end{bmatrix} \right)^T \begin{pmatrix} \sigma^2_{Y_1} & \sigma_{Y_1Y_2} \\ \sigma_{Y_2Y_1} & \sigma^2_{Y_2} \end{pmatrix}^{-1} \left(\begin{bmatrix} Y_{1i} \\ Y_{2i} \end{bmatrix} - \begin{bmatrix} \mu_{Y_1} \\ \mu_{Y_2} \end{bmatrix} \right)$$

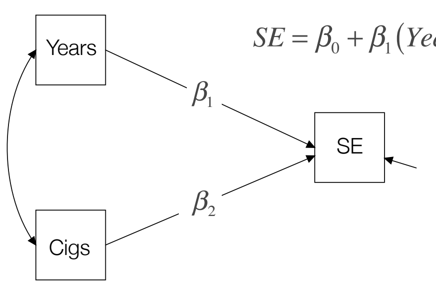
Incomplete cases

$$z_i^2 = (\mathbf{y}_i - \mu_i)^T \Sigma_i^{-1} (\mathbf{y}_i - \mu_i) = (Y_{1i} - \mu_{Y_1})^T (\sigma^2_{Y_1})^{-1} (Y_{1i} - \mu_{Y_1}) = \frac{(Y_{1i} - \mu_{Y_1})^2}{\sigma^2_{Y_1}}$$

Analysis Model

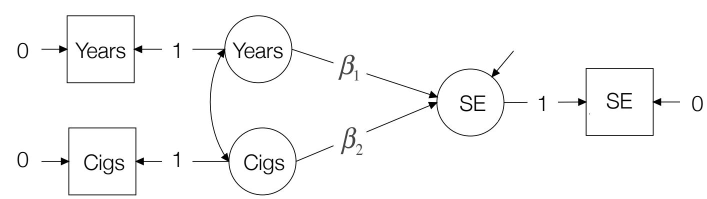
The analysis model is a multiple regression predicting self-efficacy to quit based on years smoking and number of cigarettes smoked

$$SE = \beta_0 + \beta_1(Years) + \beta_2(Cigs) + \varepsilon$$



Structural Equation Modeling Representation of a Regression Model

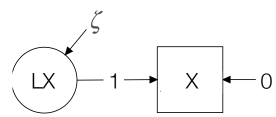
Normally distributed pseudo-latent variables share a one-to-one linkage with the manifest variables



Parameter Constraints

Loadings = 1, residual variance = 0, intercept = 0

These constraints produce a normally distributed pseudo-latent variable with the same mean and variance as its indicator



The pseudo-latent variable is a carbon copy of the indicator

Mahalanobis Distance Expression

$$z^2 = (\mathbf{y}_i - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)$$

$$= \left(\begin{bmatrix} Y_i \\ X_{1i} \\ X_{2i} \end{bmatrix} - \begin{bmatrix} \beta_0 + \beta_1 \kappa_{X_1} + \beta_2 \kappa_{X_2} \\ \kappa_{X_1} \\ \kappa_{X_2} \end{bmatrix} \right)^T \Sigma^{-1} \left(\begin{bmatrix} Y_i \\ X_{1i} \\ X_{2i} \end{bmatrix} - \begin{bmatrix} \beta_0 + \beta_1 \kappa_{X_1} + \beta_2 \kappa_{X_2} \\ \kappa_{X_1} \\ \kappa_{X_2} \end{bmatrix} \right)$$

Latent means

Latent residual variance

$$\Sigma = \begin{pmatrix} \beta\Phi\beta^T + \psi & \beta\Phi \\ \Phi\beta^T & \Phi \end{pmatrix}$$

Covariance matrix of latent predictors

Mplus Script Ex0b.inp
Maximum Likelihood Analysis

```
DATA:
file = smoking.dat;
VARIABLE:
names = id txgroup txdum1 txdum2 male age years
  cigs heavycig efficacy stress;
usevariables = years cigs efficacy;
missing = all(-99);
MODEL:
efficacy years cigs;
efficacy on years (b1)
  cigs (b2);
MODEL TEST:
b1 = 0; b2 = 0;
OUTPUT:
standardized(stdyx);
```


Analysis Comparison

Multiple Imputation

	Estimate	S.E.	Est./S.E.	P-Value
EFFICACY ON				
YEARS	-0.652	0.269	-2.423	0.015
CIGS	-0.069	0.278	-0.249	0.803
Intercepts				
EFFICACY	17.815	2.818	6.323	0.000

Maximum Likelihood

	Estimate	S.E.	Est./S.E.	P-Value
EFFICACY ON				
YEARS	-0.637	0.254	-2.505	0.012
CIGS	-0.106	0.267	-0.397	0.691
Intercepts				
EFFICACY	18.207	2.783	6.541	0.000

Conclusions

With normally distributed variables, multiple imputation and maximum likelihood estimation tend to give similar results

Maximum likelihood is preferable based on ease of use

Multiple imputation is arguably more flexible for handling complexities that arise with behavioral science data

Practical Issues: Advantage Imputation

Mixtures of categorical and continuous variables

Composite scores with missing components

Interactive (moderation) effects

Multilevel data

Practical Issue 1
Mixtures of Categorical and Continuous variables

Mixtures of Categorical and Continuous Variables

Maximum likelihood has limited capacity for handling categorical variables, particularly categorical predictors where software programs generally assume normality

Multiple imputation is ideally suited for this situation because it can tailor each variable’s regression model to match its scale

Software programs use logistic or probit regression models for categorical imputation

Complete Categorical Variables

Complete categorical variables can serve as predictors in the imputation model

Nominal variables must appear as dummy codes (Blimp's NOMINAL command automatically performs the coding)

Ordinal variables can be left as is or dummy coded

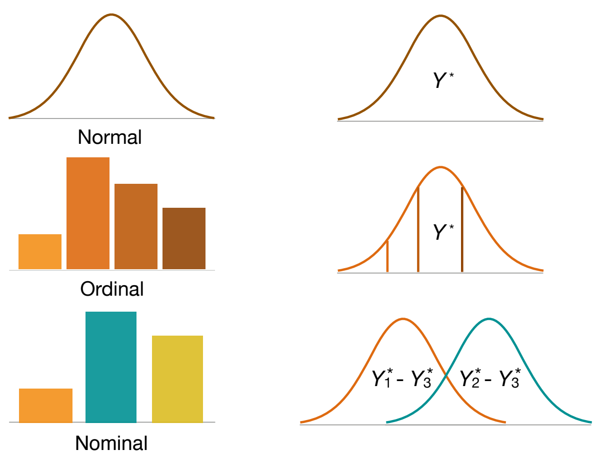
Latent Variable Formulation

Blimp uses probit regression for categorical imputation

Discrete responses arise from one or more underlying normal latent variables (Y^* variables)

Imputations are generated on the latent variable metric and are subsequently converted to discrete imputes

Latent Variable Transformations



Motivating Example

Data from a cluster-randomized study investigating a novel math problem-solving curriculum

29 schools were randomly assigned to an intervention or control condition, with an average of 33.86 students per school

We will ignore the multilevel data structure for now

Problem-Solving Data Set

Variable	Description	Missing	Metric
school	School identifier variable		
condition	Treatment code (0 = control, 1 = intervention)		Nominal
esolpercent	Percentage of English as second language	*	Numeric
student	Student identifier		
abilitygrp	Ability grouping (3-group classification)	*	Nominal
female	Female dummy code		Nominal
stanmath	Standardized math test scores	*	Numeric
frlunch	Lunch assistance dummy code	*	Nominal
efficacy	Math self-efficacy rating scale	*	Ordinal
probsolve1	Math problem-solving score at baseline	*	Numeric
probsolve7	Math problem-solving score at final wave	*	Ordinal

Analysis Model

The substantive analysis is a regression model where a number of student-level covariates predict end-of-year problem-solving

$$probsolve7 = \beta_0 + \beta_1(probsolve1) + \beta_2(efficacy) + \beta_3(female) + \beta_4(abilitygrp2) + \beta_5(abilitygrp3) + \varepsilon$$

The ordinal self-efficacy ratings and nominal ability grouping variables are incomplete

Blimp Script Ex1a.imp Diagnostic Phase

```
DATA: ~/desktop/examples/probsolve.dat;
VARIABLES: school condition esolpercent student abilitygrp female
   stanmath frlunch efficacy probsolve1 probsolve7;
ORDINAL: efficacy;
NOMINAL: abilitygrp female frlunch;
MISSING: -99;
MODEL: ~ abilitygrp female stanmath frlunch efficacy
   probsolve1 probsolve7;
NIMPS: 2;
BURN: 3000;
THIN: 1;
SEED: 90291;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate psr;
CHAINS: 2 processors 2;
```


Diagnostic Output

```
Comparing iterations 751 to 1500 for 2 chains.
                    -----------------------------------------------------
                    |     Fix Eff| Ran Eff Var|     Err Var|   Threshold|
                    -----------------------------------------------------
           Max PSR  |       1.054|         nan|       1.002|       1.078|
  Missing Variable  |  abilitygrp|            |  probsolve7|    efficacy|
                    -----------------------------------------------------

Comparing iterations 801 to 1600 for 2 chains.
                    -----------------------------------------------------
                    |     Fix Eff| Ran Eff Var|     Err Var|   Threshold|
                    -----------------------------------------------------
           Max PSR  |       1.050|         nan|       1.003|       1.034|
  Missing Variable  |  abilitygrp|            |  probsolve7|    efficacy|
                    -----------------------------------------------------
```

Blimp Script Ex1b.imp
Imputation Phase (Mplus Format)

```
DATA: ~/desktop/examples/probsolve.dat;
VARIABLES: school condition esolpercent student abilitygrp female
   stanmath frlunch efficacy probsolve1 probsolve7;
ORDINAL: efficacy;
NOMINAL: abilitygrp female frlunch;
MISSING: -99;
MODEL: ~ abilitygrp female stanmath frlunch efficacy
   probsolve1 probsolve7;
NIMPS: 20;
BURN: 1000;
THIN: 1000;
SEED: 90291;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate;
CHAINS: 2 processors 2;
```


Blimp Script Ex1c.imp
Imputation Phase (R, SAS, SPSS, and Stata Format)

```
DATA: ~/desktop/examples/probsolve.dat;
VARIABLES: school condition esolpercent student abilitygrp female
   stanmath frlunch efficacy probsolve1 probsolve7;
ORDINAL: efficacy;
NOMINAL: abilitygrp female frlunch;
MISSING: -99;
MODEL: ~ abilitygrp female stanmath frlunch efficacy
   probsolve1 probsolve7;
NIMPS: 20;
BURN: 1000;
THIN: 1000;
SEED: 90291;
OUTFILE: ~/desktop/examples/imps.csv;
OPTIONS: stacked;
CHAINS: 2 processors 2;
```


Mplus Script Ex1.inp
Analysis and Pooling Phase

```
DATA:
file = implist.csv;
type = imputation;
VARIABLE:
names =  school condition esolpercent student abilgrp female
   stanmath frlunch efficacy probsolve1 probsolve7;
usevariables = female efficacy probsolve1 probsolve7
   abilgrp2 abilgrp3;
DEFINE:
abilgrp2 = 0;
abilgrp3 = 0;
if (abilgrp eq 2) then abilgrp2 = 1;
if (abilgrp eq 3) then abilgrp3 = 1;
MODEL:
probsolve7 on probsolve1 efficacy female abilgrp2 abilgrp3;
OUTPUT:
standardized;
```


Mplus Output

```
MODEL RESULTS
                                                    Two-Tailed
                     Estimate       S.E.  Est./S.E.    P-Value
 PROBSOLV ON
    PROBSOLVE1          0.444      0.044     10.072      0.000
    EFFICACY            0.732      0.297      2.464      0.014
    FEMALE              0.146      0.764      0.191      0.849
    ABILGRP2            0.470      1.445      0.325      0.745
    ABILGRP3            3.854      1.650      2.335      0.020
 Intercepts
    PROBSOLVE7         60.034      4.445     13.505      0.000
 Residual Variances
    PROBSOLVE7        114.089      5.717     19.956      0.000
```

Practical Issue 2
Composite Scores with Missing Components

Scale Scores

Measuring complex psychological constructs requires multiple questionnaire items, each of which taps into a different aspect of the construct

Researchers often compute scale scores by summing or averaging questionnaire items

How to compute the composite when its constituent items are incomplete?

Prorated Scale Scores (Averaging the Available Items)

Researchers often compute so-called prorated scale scores by averaging the available item responses

e.g., A respondent who answered 7 out of 10 items has a scale score equal to the average of the 7 responses

This approach is not ideal because it makes stringent assumptions that are unlikely to hold in practice

Proration = Person Mean Imputation

Prorated Scale Scores

ID	Q1	Q2	Q3	Scale
1	1	2	1	1.3
2	5	NA	4	4.5
3	3	2	4	3.0
4	NA	3	NA	3.0

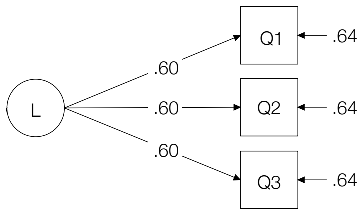
Person-Mean Imputation

ID	Q1	Q2	Q3	Scale
1	1	2	1	1.3
2	5	4.5	4	4.5
3	3	2	4	3.0
4	3.0	3	3.0	3.0

Proration Assumptions

Proration requires MCAR plus identical item means and inter-item correlations (parallel factor structure)

	Q1	Q2	Q3
Q1	1.00		
Q2	0.36	1.00	
Q3	0.36	0.36	1.00
Means	3.00	3.00	3.00



Scale-Level vs. Item-Level Missing Data Handling

Scale-level imputation fills in the incomplete composite scores, ignoring item-level information

Item-level imputation fills in the items, and the composite is subsequently computed during the analysis phase

Item-level imputation offers a dramatic gain in power

Scale-Level Imputation

Component Variables

ID	X_1	X_2	X_3	Y_1	Y_2
1	1	2	1	NA	3
2	5	NA	4	NA	NA
3	3	2	4	3	4
4	NA	3	NA	5	5
…					
200	4	5	4	3	4

Imputation Variables

ID	Scale X	Scale Y
1	4	NA
2	NA	NA
3	9	7
4	NA	10
…		
200	13	7

Item-Level Imputation

Component Variables

ID	X_1	X_2	X_3	Y_1	Y_2
1	1	2	1	NA	3
2	5	NA	4	NA	NA
3	3	2	4	3	4
4	NA	3	NA	5	5
…					
200	4	5	4	3	4

Imputation Variables

ID	X_1	X_2	X_3	Y_1	Y_2
1	1	2	1	NA	3
2	5	NA	4	NA	NA
3	3	2	4	3	4
4	NA	3	NA	5	5
…					
200	4	5	4	3	4

Motivating Example

Questionnaire data from a study of eating disorder risk in a sample of 500 college-aged women

Variables include body mass index (BMI), questionnaire items measuring body dissatisfaction and eating disorder risk, past sexual abuse history (0 = no abuse history, 1 = abuse history)

All questionnaire items measured on a 7-point scale

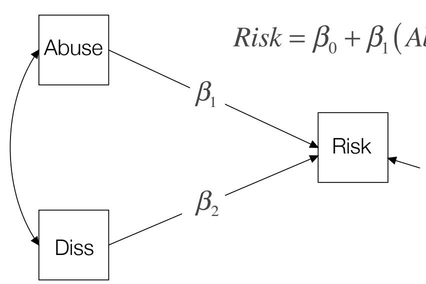
Eating Disorder Risk Data

Variable	Description	Missing	Metric
abuse	Previous history of abuse indicator	*	Nominal
bmi	Body mass index	*	Numeric
bds1 - bds7	Body dissatisfaction questionnaire items	*	Ordinal
edr1 - edr6	Eating disorder risk questionnaire items	*	Ordinal

Analysis Model

Body dissatisfaction and eating disorder risk are scale scores computed as the sum of the item responses

$$Risk = \beta_0 + \beta_1 (Abuse) + \beta_2 (Diss) + \varepsilon$$



Blimp Script Ex2a.imp Diagnostic Phase

```
DATA: ~/desktop/examples/eatingrisk.dat;
VARNAMES: abuse bmi bds1-bds7 edr1-edr6;
NOMINAL: abuse;
ORDINAL: bds1-bds7 edr1-edr6;
MISSING: -99;
MODEL: ~ abuse bmi bds1-bds7 edr1-edr6;
SEED: 90291;
BURN: 20000;
THIN: 1;
NIMPS: 2;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate psr;
CHAINS: 2 processors 2;
```

Diagnostic Output

```
Comparing iterations 9251 to 18500 for 2 chains.
                  -----------------------------------------------------
                  |    Fix Eff| Ran Eff Var|    Err Var|   Threshold|
                  -----------------------------------------------------
         Max PSR  |      1.007|         nan|      1.000|       1.055|
Missing Variable  |       edr2|            |        bmi|        bds1|
                  -----------------------------------------------------

Comparing iterations 9301 to 18600 for 2 chains.
                  -----------------------------------------------------
                  |    Fix Eff| Ran Eff Var|    Err Var|   Threshold|
                  -----------------------------------------------------
         Max PSR  |      1.007|         nan|      1.000|       1.046|
Missing Variable  |       edr2|            |        bmi|        bds1|
                  -----------------------------------------------------
```


Blimp Script Ex2b.imp
Imputation Phase (Mplus Format)

```
DATA: ~/desktop/examples/eatingrisk.dat;
VARNAMES: abuse bmi bds1-bds7 edr1-edr6;
NOMINAL: abuse;
ORDINAL: bds1-bds7 edr1-edr6;
MISSING: -99;
MODEL: ~ abuse bmi bds1-bds7 edr1-edr6;
SEED: 90291;
BURN: 10000;
THIN: 10000;
NIMPS: 20;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate;
CHAINS: 2 processors 2;
```


Blimp Script Ex2c.imp
Imputation Phase (R, SAS, SPSS, and Stata Format)

```
DATA: ~/desktop/examples/eatingrisk.dat;
VARNAMES: abuse bmi bds1-bds7 edr1-edr6;
NOMINAL: abuse;
ORDINAL: bds1-bds7 edr1-edr6;
MISSING: -99;
MODEL: ~ abuse bmi bds1-bds7 edr1-edr6;
SEED: 90291;
BURN: 10000;
THIN: 10000;
NIMPS: 20;
OUTFILE: ~/desktop/examples/imps.csv;
OPTIONS: stacked;
CHAINS: 2 processors 2;
```


Mplus Script Ex2a.inp
Analysis and Pooling Phase

```
DATA:
file = implist.csv;
type = imputation;
VARIABLE:
names = abuse bmi bds1-bds7 edr1-edr6;
usevariables = abuse bodydis eatrisk;
DEFINE:
bodydis = sum(bds1-bds7);
eatrisk = sum(edr1-edr6);
MODEL:
eatrisk on abuse bodydis;
OUTPUT:
standardized(stdyx);
```

Mplus Output

```
MODEL RESULTS
                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value
 EATRISK   ON
    ABUSE              1.355      0.512      2.646      0.008
    BODYDIS            0.500      0.031     16.188      0.000
 Intercepts
    EATRISK           10.092      0.879     11.479      0.000
 Residual Variances
    EATRISK           12.142      0.800     15.170      0.000
```


Mplus Output, Continued

```
STANDARDIZED MODEL RESULTS

STDYX Standardization
                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value
 EATRISK   ON
    ABUSE              0.108      0.041      2.648      0.008
    BODYDIS            0.601      0.031     19.441      0.000
 Intercepts
    EATRISK            2.211      0.243      9.109      0.000
 Residual Variances
    EATRISK            0.583      0.035     16.815      0.000

R-SQUARE
    Observed                                        Two-Tailed
    Variable        Estimate       S.E.  Est./S.E.    P-Value
    EATRISK            0.417      0.035     12.034      0.000
```


Mplus Script Ex2b.inp
Scale-Level Maximum Likelihood Analysis

```
DATA:
file = eatingrisk.dat;
VARIABLE:
names = abuse bmi bds1-bds7 edr1-edr6;
usevariables = abuse bodydis eatrisk;
missing = all(-99);
DEFINE:
bodydis = sum(bds1-bds7);
eatrisk = sum(edr1-edr6);
MODEL:
eatrisk abuse bodydis
eatrisk on abuse bodydis;
OUTPUT:
standardized(stdyx);
```


Analysis Comparison

Item-Level Multiple Imputation

```
                    Estimate       S.E.  Est./S.E.    P-Value
EATRISK   ON
   ABUSE               1.355      0.512      2.646      0.008
   BODYDIS             0.500      0.031     16.188      0.000
 Intercepts
   EATRISK            10.092      0.879     11.479      0.000
```

Scale-Level Maximum Likelihood

```
                    Estimate       S.E.  Est./S.E.    P-Value
 EATRISK   ON
   ABUSE               1.837      0.664      2.768      0.006
   BODYDIS             0.514      0.041     12.507      0.000
 Intercepts
   EATRISK             9.683      1.131      8.564      0.000
```

Important Conclusions

Item-level imputation offers a dramatic gain in precision

The scale-level analysis would require a 60% increase in sample size to achieve the same standard errors as item-level missing data handling

e.g., Reducing the abuse coefficient's standard error from .664 to .512 requires an increase from N = 500 to 790

Practical Issue 3
Incomplete Interaction Effects

Interaction (Moderation)

Moderation (interaction) occurs when the magnitude of a bivariate relation depends on a third variable

In a regression analysis, the influence of the focal predictor depends on the value of the moderator

e.g., The influence of pain severity (focal) on daily stress (outcome) is different for males and females (moderator)

Just-Another-Variable Imputation

The just-another-variable method treats the interaction as missing when one of its is missing, and it imputes the product like any other normally distributed variable

Product terms cannot follow a normal distribution, and imputing an interaction can introduce substantial bias unless the mechanism is MCAR

Maximum likelihood suffers from the same problem

Substantive Model-Compatible Imputation

Substantive model-compatible imputation does not impute the product or specify its distribution

The interaction components are imputed from a model that includes only other predictors (no product)

A special algorithm (Metropolis) selects imputations that are consistent with a moderated regression

The outcome variable is imputed from a model that includes the product of the imputed predictor variables

Motivating Example

Diary data from a sample of 250 chronic pain patients

Variables include gender, number of diagnosed medical conditions, sleep quality ratings, and scale scores measuring pain severity, positive affect, negative affect, and daily life stress

Pain Data

Variable	Description	Missing	Metric
female	Gender dummy code		Nominal
diagnose	Number of diagnosed medical problems		Count
sleep	Likert-type sleep quality rating	*	Ordinal
pain	Pain severity scale score	*	Numeric
posaff	Positive affect scale score	*	Numeric
negaff	Negative affect scale score	*	Numeric
stress	Stress scale score	*	Numeric

Analysis Example

The analysis is a regression that examines whether the influence of pain on stress differs for males and females

$$stress = \beta_0 + \beta_1(pain) + \beta_2(female) + \beta_3(pain)(female) + \varepsilon$$

Stress and pain (and thus the product) are incomplete

Blimp Script Ex3a.imp Diagnostic Phase

```
DATA: ~/desktop/examples/pain.dat;
VARNAMES: female diagnose sleep pain posaff negaff stress;
MISSING: -99;
MODEL: ~ female diagnose sleep pain negaff stress pain*female;
ORDINAL: sleep;
OUTCOME: stress;
SEED: 90291;
BURN: 3000;
THIN: 1;
NIMPS: 2;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate psr;
CHAINS: 2 processors 2;
```


Diagnostic Output

```
Comparing iterations 401 to 800 for 2 chains.
                    -----------------------------------------------------
                    |     Fix Eff| Ran Eff Var|     Err Var|   Threshold|
                    -----------------------------------------------------
           Max PSR  |       1.019|         nan|       1.005|       1.089|
  Missing Variable  |       sleep|            |        pain|       sleep|
                    -----------------------------------------------------

Comparing iterations 451 to 900 for 2 chains.
                    -----------------------------------------------------
                    |     Fix Eff| Ran Eff Var|     Err Var|   Threshold|
                    -----------------------------------------------------
           Max PSR  |       1.010|         nan|       1.003|       1.030|
  Missing Variable  |      stress|            |        pain|       sleep|
                    -----------------------------------------------------
```


Blimp Script Ex3b.imp Imputation Phase (Mplus Format)

```
DATA: ~/desktop/examples/pain.dat;
VARNAMES: female diagnose sleep pain posaff negaff stress;
MISSING: -99;
MODEL: ~ female diagnose sleep pain negaff stress pain*female;
ORDINAL: sleep;
OUTCOME: stress;
SEED: 90291;
BURN: 500;
THIN: 500;
NIMPS: 20;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate;
CHAINS: 2 processors 2;
```


Blimp Script Ex3c.imp Imputation Phase (R, SAS, SPSS, and Stata Format)

```
DATA: ~/desktop/examples/pain.dat;
VARNAMES: female diagnose sleep pain posaff negaff stress;
MISSING: -99;
MODEL: ~ female diagnose sleep pain negaff stress pain*female;
ORDINAL: sleep;
OUTCOME: stress;
SEED: 90291;
BURN: 500;
THIN: 500;
NIMPS: 20;
OUTFILE: ~/desktop/examples/imps.csv;
OPTIONS: stacked;
CHAINS: 2 processors 2;
```

Mplus Script Ex3.inp
Analysis and Pooling Phase

```
DATA:
file = implist.csv;
type = imputation;
VARIABLE:
names = female diagnose sleep pain posaff negaff stress;
usevariables = stress female pain femxpain;
DEFINE:
femxpain = female*pain;
MODEL:
stress on female pain femxpain;
OUTPUT:
standardized;
```


Mplus Output

```
MODEL RESULTS
                                                        Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value
 STRESS   ON
    FEMALE            -1.508      0.655     -2.302      0.021
    PAIN               0.141      0.094      1.495      0.135
    FEMXPAIN           0.296      0.138      2.147      0.032
 Intercepts
    STRESS             3.179      0.395      8.045      0.000
 Residual Variances
    STRESS             0.762      0.081      9.356      0.000
```


Practical Issue 4
Multilevel Data Structures

Multilevel Data

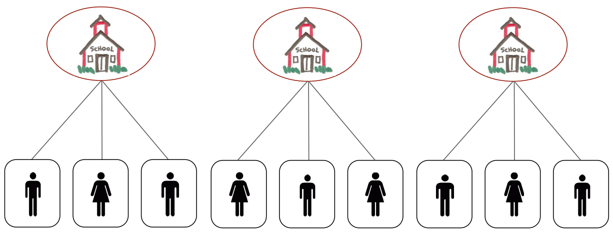
A unit of analysis is the what or whom being studied (e.g., observations, individuals, classrooms, groups, families, etc.)

Multilevel data structures have multiple units of analysis that are hierarchically nested

Lower-level units are nested within higher-level units

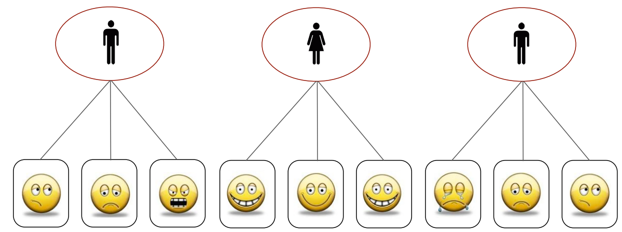
Multilevel Data Example 1

Sample comprised of multiple schools and several students in each school (i.e., students nested within schools)



Multilevel Data Example 2

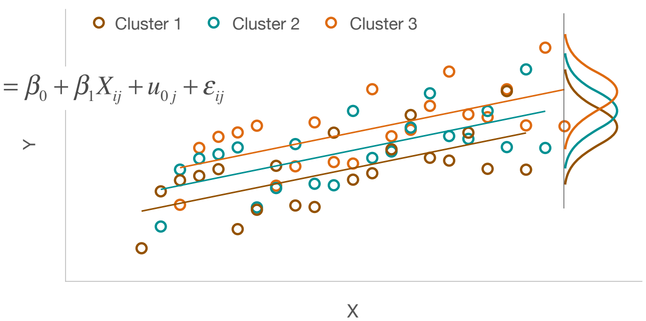
Sample comprised of multiple individuals, each with several daily assessments of mood (i.e., observations nested within individuals)



Random Intercept Model

A random intercept model is one where only the means vary across clusters

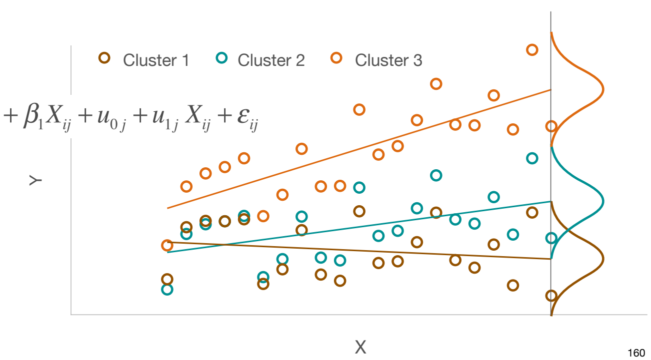
$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_{0j} + \varepsilon_{ij}$$



Random Slope Model

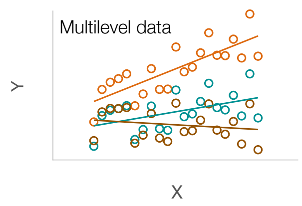
A random slope model allows the relation between a pair of level-1 variables to differ across clusters

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_{0j} + u_{1j} X_{ij} + \varepsilon_{ij}$$

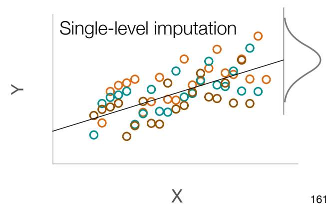


Single-Level Imputation

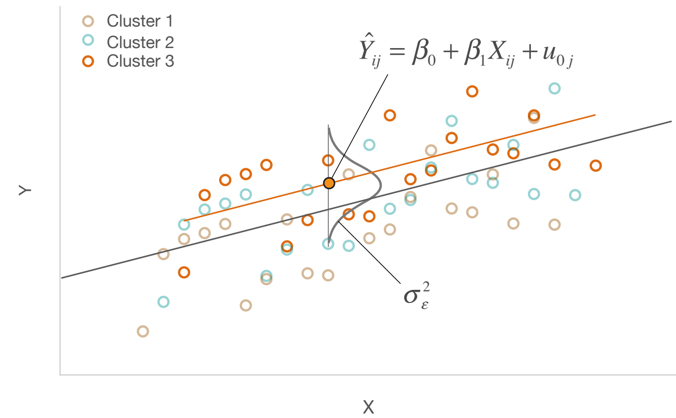
Standard imputation routines assume a common distribution for all clusters (same means and variance-covariance matrix)



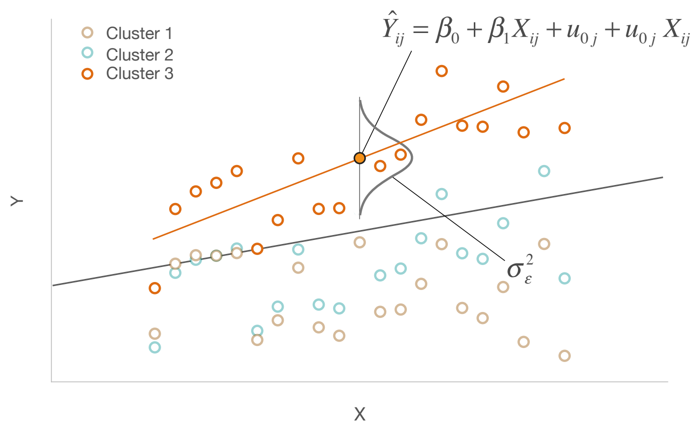
Imputation will introduce substantial bias under any mechanism



Random Intercept Imputation Model



Random Slope Imputation Model



Motivating Example

Data from a cluster-randomized study investigating a novel math problem-solving curriculum

29 schools (level-2 units) were randomly assigned to an intervention or control condition

The average number of students (level-1 units) per school was 33.86, with a range of 13 to 61

School Study Data

	Variable	Description	Missing	Metric
Level-2	school	School identifier variable		
	condition	Treatment code (0 = control, 1 = intervention)		Nominal
	esolpercent	Percentage of English as second language	*	Numeric
Level-1	student	Student identifier		
	abilitygrp	Ability grouping (3-group classification)	*	Nominal
	female	Female dummy code		Nominal
	stanmath	Standardized math test scores	*	Numeric
	frlunch	Lunch assistance dummy code	*	Nominal
	efficacy	Math self-efficacy rating scale	*	Ordinal
	probsolve1	Math problem-solving score at baseline	*	Numeric
	probsolve7	Math problem-solving score at final wave	*	Ordinal

Analysis Model

The substantive analysis is a random slope model where intervention condition and covariates predict end-of-year problem-solving scores, with self-efficacy ratings as a random predictor

$$\begin{aligned} probsolve7_{ij} &= \beta_0 + \beta_1(probsolve1_{ij}) + \beta_2(efficacy_{ij}) + \beta_3(female_{ij}) \\ &+ \beta_4(esolpercent_j) + \beta_5(condition_j) \\ &+ u_{0j} + u_{1j}(efficacy_{ij}) + \varepsilon \end{aligned}$$

Substantive Model-Compatible Imputation

Fully conditional specification uses a “reverse random coefficient” approach that negatively biases variance estimates when predictors are missing

Substantive model-compatible imputation is better suited for models with random slopes

Same idea as an interaction, as a random slope is just the product of a latent variable (u_{1j}) and manifest variable

Blimp Script Ex4a.imp Diagnostic Phase

```
DATA: ~/desktop/examples/probsolve.dat;
VARIABLES: school condition esolpercent student abilitygrp
   female stanmath frlunch efficacy probsolve1 probsolve7;
ORDINAL: condition female frlunch efficacy;
OUTCOME: probsolve7;
MISSING: -99;
MODEL: school ~ condition esolpercent female stanmath
   frlunch probsolve1 efficacy:probsolve7;
NIMPS: 2;
BURN: 3000;
THIN: 1;
SEED: 90291;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate psr;
CHAINS: 2 processors 2;
```

Diagnostic Output

```
Comparing iterations 401 to 800 for 2 chains.
                    -------------------------------------------------------
                    |      Fix Eff| Ran Eff Var|      Err Var|    Threshold|
                    -------------------------------------------------------
         Max PSR    |        1.051|        1.015|        1.005|          nan|
 Missing Variable   |  probsolve7|  probsolve7|  probsolve7|             |
                    -------------------------------------------------------

Comparing iterations 451 to 900 for 2 chains.
                    -------------------------------------------------------
                    |      Fix Eff| Ran Eff Var|      Err Var|    Threshold|
                    -------------------------------------------------------
         Max PSR    |        1.029|        1.025|        1.002|          nan|
 Missing Variable   |  probsolve7|  probsolve7|  probsolve7|             |
                    -------------------------------------------------------
```


Blimp Script Ex4b.imp
Imputation Phase (Mplus Format)

```
DATA: ~/desktop/examples/probsolve.dat;
VARIABLES: school condition esolpercent student abilitygrp
   female stanmath frlunch efficacy probsolve1 probsolve7;
ORDINAL: condition female frlunch efficacy;
OUTCOME: probsolve7;
MISSING: -99;
MODEL: school ~ condition esolpercent female stanmath
   frlunch probsolve1 efficacy:probsolve7;
NIMPS: 20;
BURN: 500;
THIN: 500;
SEED: 90291;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate;
CHAINS: 2 processors 2;
```


Blimp Script Ex4c.imp
Imputation Phase ((R, SAS, SPSS, and Stata Format)

```
DATA: ~/desktop/examples/probsolve.dat;
VARIABLES: school condition esolpercent student abilitygrp
   female stanmath frlunch efficacy probsolve1 probsolve7;
ORDINAL: condition female frlunch efficacy;
OUTCOME: probsolve7;
MISSING: -99;
MODEL: school ~ condition esolpercent female stanmath
   frlunch probsolve1 efficacy:probsolve7;
NIMPS: 20;
BURN: 500;
THIN: 500;
SEED: 90291;
OUTFILE: ~/desktop/examples/imps.csv;
OPTIONS: stacked;
CHAINS: 2 processors 2;
```


Mplus Script Ex4.inp
Analysis and Pooling Phase

```
DATA:
file = implist.csv;
type = imputation;
VARIABLE:
names =  school condition esolpercent student abilgrp female stanmath
   frlunch efficacy probsolve1 probsolve7;
usevariables = condition esolpercent female efficacy probsolve1 probsolve7;
cluster = school;
within = female efficacy probsolve1;
between = condition esolpercent;
ANALYSIS:
type = twolevel random;
MODEL:
%within%
ranslope | probsolve7 on efficacy;
probsolve7 on probsolve1 female;
%between%
probsolve7 on esolpercent condition;
probsolve7 with ranslope;
```

Mplus Output

```
MODEL RESULTS
                                                   Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value
Within Level
 PROBSOLVE7 ON
    PROBSOLVE1         0.437      0.036     12.260      0.000
    FEMALE             0.319      0.683      0.467      0.641
 Residual Variances
    PROBSOLVE7        89.677      6.201     14.463      0.000
Between Level
 PROBSOLVE7 ON
    ESOLPERCEN         0.078      0.039      1.981      0.048
    CONDITION          5.001      1.825      2.740      0.006
```


Mplus Output, Continued

```
MODEL RESULTS
                                                   Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value
 PROBSOLV WITH
    RANSLOPE          -1.339      2.081     -0.643      0.520
 Means
    RANSLOPE           0.824      0.271      3.040      0.002
 Intercepts
    PROBSOLVE7        54.070      4.448     12.157      0.000
 Variances
    RANSLOPE           0.256      0.434      0.589      0.556
 Residual Variances
    PROBSOLVE7        24.338     10.724      2.270      0.023
```