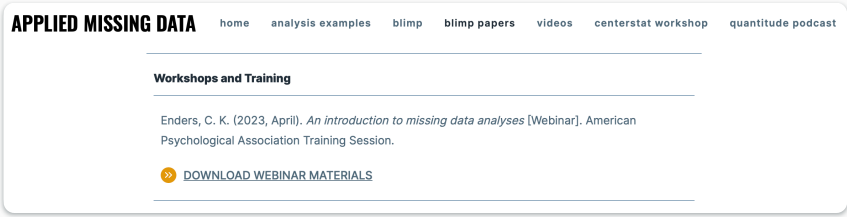
APA SCIENCE TRAINING SESSION

AN INTRODUCTION TO MISSING DATA ANALYSES

Craig K. Enders
UCLA Department of Psychology

1

WWW.APPLIEDMISSINGDATA.COM/BLIMP-PAPERS



2

AGENDA

1. Modern Missing Data Methods
2. Missing Data Mechanisms
3. Maximum Likelihood Estimation
4. Bayesian Estimation
5. Multiple Imputation
6. Missing Data Software
7. Analysis Example

MODERN MISSING DATA METHODS

MODERN MISSING DATA METHODS

Maximum likelihood

Bayesian estimation

Multiple imputation

the Big Three

KEY ADVANTAGES OF BIG THREE

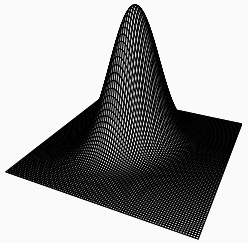
- Achieve unbiasedness with more a realistic assumption about the missing data process
- Allow for alternate assumptions about nonresponse process
- Maximize power
- Use all available data, no wasted resources

CHOOSING A MISSING DATA METHOD

- All things being equal—same data, same variables, same assumptions—the Big Three rarely produce different results
- Missing data analyses require distributional assumptions
- How we represent those distributions—multivariate versus factored specifications—is what matters

MODELING FRAMEWORKS

Multivariate modeling

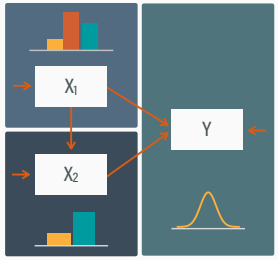


- Classic approaches often assume multivariate normality
- Most applications of maximum likelihood and multiple imputation

MODELING FRAMEWORKS

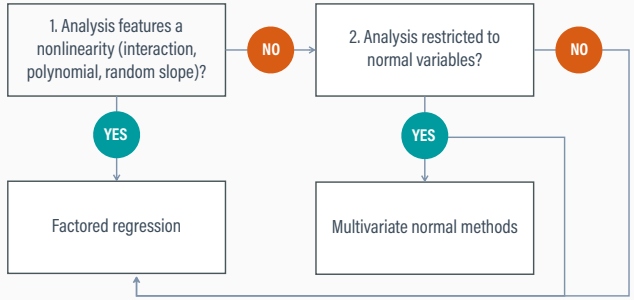
Multivariate modeling

Factored regression specification



- Factored regression invokes a unique model and distribution for each variable
- Each model can include terms that are at odds with multivariate normality (e.g., interactions, random slopes)

MISSING DATA DECISION TREE



MISSING DATA MECHANISMS

HOW MUCH MISSING DATA IS TOO MUCH?

- The Big Three can tolerate substantial amounts of missing data
- The Big Three are increasingly better than ad hoc methods (e.g., deleting incomplete cases) as missingness increases
- The amount of missing data is less important than why the data are missing (the missingness process or mechanism)

MISSING DATA MECHANISMS

- Missing data mechanisms (processes) describe different ways in which the data relate to nonresponse
- Missingness may be completely random or systematically related to different parts of the data
- Mechanisms function as statistical assumptions

PARTITIONING THE DATA

Complete = Observed + Missing

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

=

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

+

Y_1	Y_2	Y_3
	3	
2		
	2	**2**
	5	

Indicators

M_1	M_2	M_3
0	0	0
0	**1**	0
0	0	0
1	0	0
0	0	0
0	**1**	**1**
0	0	0
0	0	0
0	**1**	0

MISSING COMPLETELY AT RANDOM

- The probability of missing values is completely unrelated to the data
- MCAR is purely random missingness
- We don't care about this process or testing for it (e.g., Little's MCAR test)

Relation between nonresponse and data

M			Y_{obs}			Y_{mis}		
M_1	M_2	M_3	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3
0	0	0	4	4	3			
0	1	0	3	NA	5		3	
0	0	0	7	1	6			
1	0	0	NA	1	6	2		
0	0	0	5	9	3			
0	1	1	3	NA	NA		2	2
0	0	0	1	6	7			
0	0	0	9	4	9			
0	1	0	2	NA	6		5	

(CONDITIONALLY) MISSING AT RANDOM

- Systematic missingness related to the observed scores
- The probability of missing values is unrelated to the unseen (latent) data
- The Big Three assume CMAR by default

Relation between nonresponse and data

M			Y_{obs}			Y_{mis}		
M_1	M_2	M_3	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3
0	0	0	4	4	3			
0	1	0	3	NA	5		3	
0	0	0	7	1	6			
1	0	0	NA	1	6	2		
0	0	0	5	9	3			
0	1	1	3	NA	NA		2	2
0	0	0	1	6	7			
0	0	0	9	4	9			
0	1	0	2	NA	6		5	

MISSING NOT AT RANDOM

- Systematic missingness
- The probability of missing values is related to the unseen (latent) data
- The Big Three can model MNAR processes (selection and pattern mixture models)

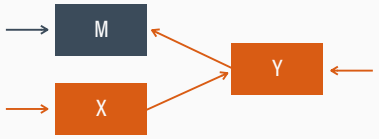
Relation between nonresponse and data

M			Y_{obs}			Y_{mis}		
M_1	M_2	M_3	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3
0	0	0	4	4	3			
0	1	0	3	NA	5		3	
0	0	0	7	1	6			
1	0	0	NA	1	6	2		
0	0	0	5	9	3			
0	1	1	3	NA	NA		2	2
0	0	0	1	6	7			
0	0	0	9	4	9			
0	1	0	2	NA	6		5	

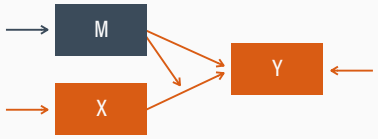
MNAR MODELING

- Missing not at random processes require an explicit model that incorporates the missing data indicator(s)

Selection Model

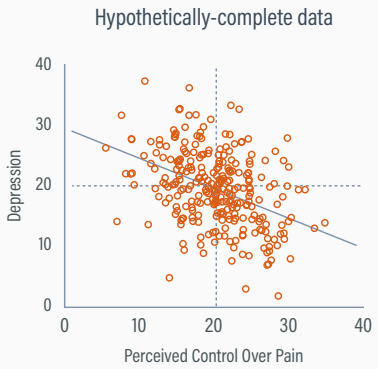


Pattern Mixture Model



CHRONIC PAIN EXAMPLE

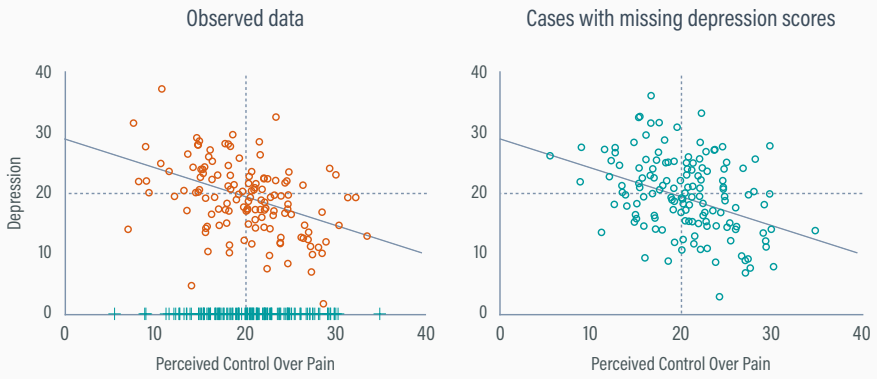
- Study investigating the psychological correlates of chronic pain
- Perceived control over pain is complete, 50% of depression scores are missing



MISSING COMPLETELY AT RANDOM

- Missingness is unrelated to the observed data (perceived control over pain) and the unseen data (depression)
- To reduce respondent burden and data collection costs, depression scores are collected from a random subset of the full sample (i.e., a planned missing data design)

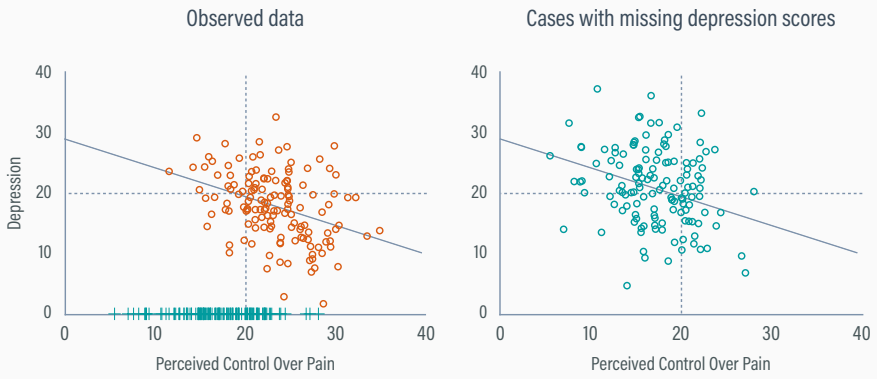
MCAR GRAPHIC



CONDITIONALLY MISSING AT RANDOM

- Missingness is related to the observed data (perceived control over pain) but unrelated to the unseen data (depression)
- Individuals with low perceived control are more likely to have missing data due to their diminished functional status
- The Big Three assume a CMAR process by default

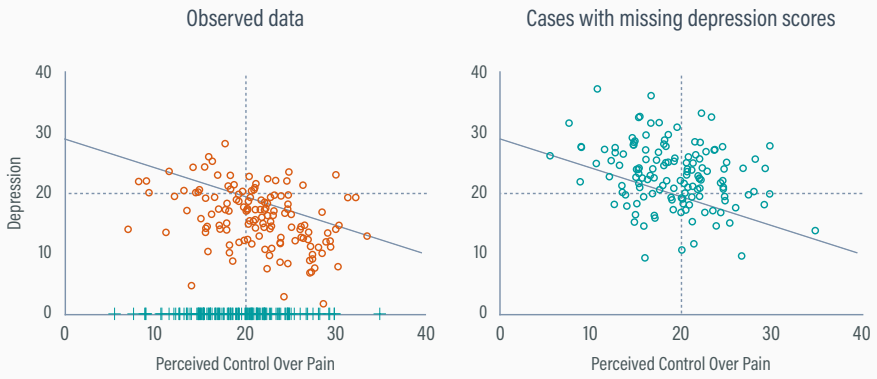
CMAR GRAPHIC



MISSING NOT AT RANDOM

- Missingness is related to the unseen data (depression) and potentially to the observed data (perceived control) as well
- Individuals with the highest levels of depression are more likely to be missing due to their debilitating symptoms

MISSING NOT AT RANDOM



TESTING THE CMAR ASSUMPTION

- The Big Three achieve unbiasedness if the process is conditionally MAR
- The CMAR assumption is untestable because it stipulates no relation between missingness and the unseen scores
- When in doubt, conduct sensitivity analyses that compare the estimates from CMAR and MNAR assumptions

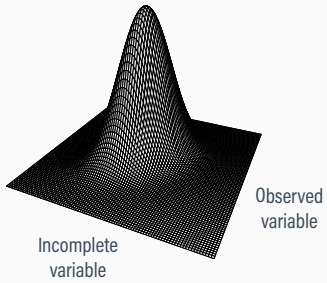
MAXIMUM LIKELIHOOD ESTIMATION

MAXIMUM LIKELIHOOD

- ML identifies parameter estimates that minimize the distances between the model's predicted values and the observed data
- Each observation's contribution to estimation is restricted to the subset of parameters for which there is data
- Estimation uses incomplete data, no imputation performed

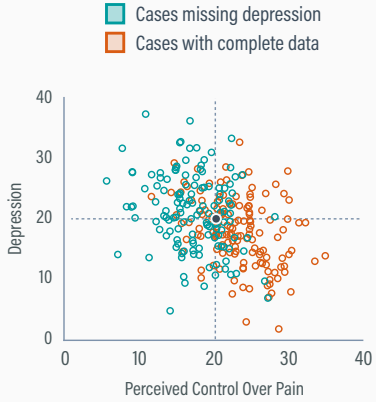
IMPLICIT IMPUTATION

- Each participant contributes their observed data
- Data are not filled in, but the multivariate normal distribution acts like an imputation machine
- The location of the observed data implies the probable position of the unseen data, and estimates are adjusted accordingly



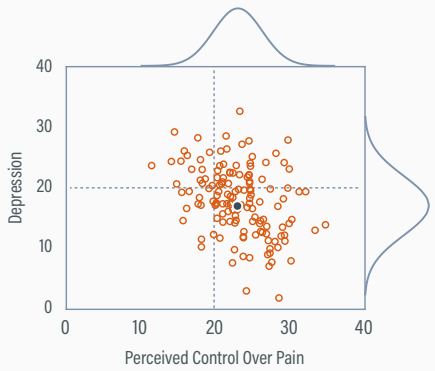
CHRONIC PAIN ILLUSTRATION

- Participants with low perceived control are more likely to have missing depression scores (conditionally MAR)
- The true means are both 20



DELETING INCOMPLETE DATA

- Deleting cases with missing depression scores gives a non-representative sample
- The perceived control mean is too high (M_{pc} = 23.1), and the depression mean is too low (M_{dep} = 17.2)

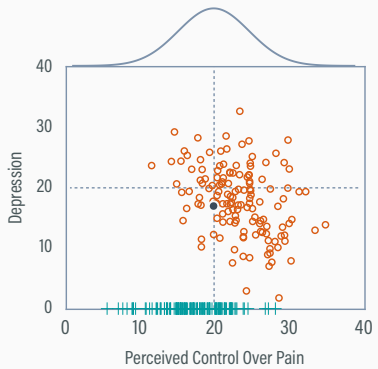


PARTIAL DATA RECORDS

- Incorporating the partial data gives a complete set of perceived control scores
- The partial data records primarily have low perceived control scores

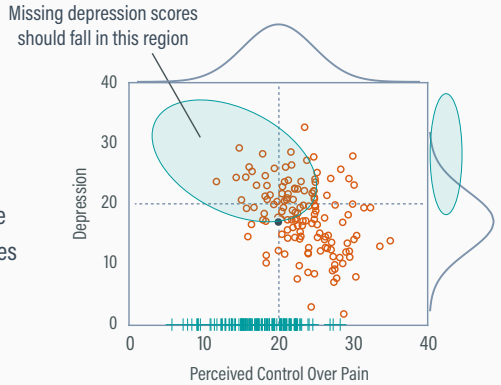
ADJUSTING PERCEIVED CONTROL MEAN

- Adding low perceived control scores increases the variable's variability
- The perceived control mean receives a downward adjustment to accommodate the influx of low scores



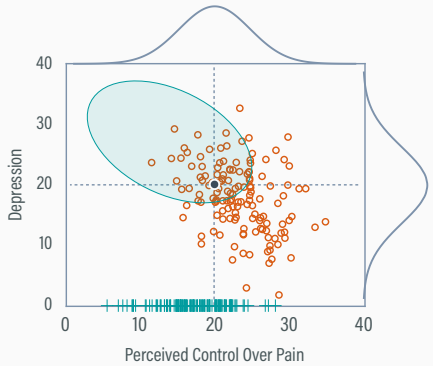
IMPLICIT IMPUTATION

- Maximum likelihood assumes multivariate normality
- In a normal distribution with a negative correlation, low perceived control scores should pair with high depression

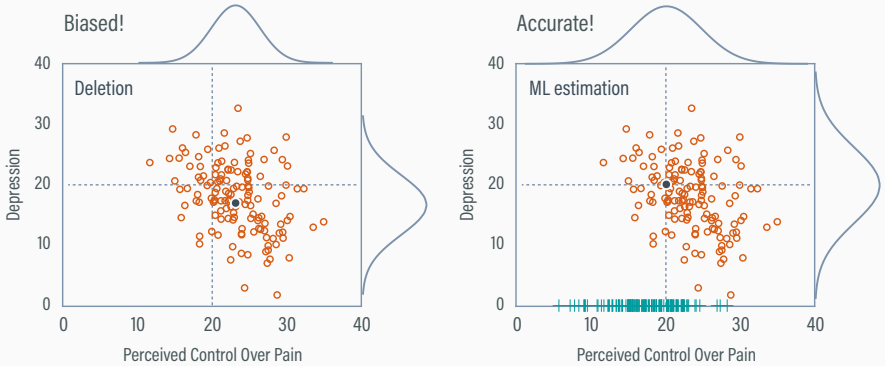


ADJUSTING THE DEPRESSION DISTRIBUTION

- Maximum likelihood intuits the presence of the elevated but unseen depression scores
- The mean and variance of depression increase to accommodate observed perceived control scores at the low end



ESTIMATE COMPARISON



MAXIMUM LIKELIHOOD PROS AND CONS

Pros	Cons
Direct estimation for a wide range of analysis models	Generally limited to normal data, options for mixed metrics are less common
Widely available in software packages (any SEM program)	Normal-theory methods are biased with interactions and non-linear terms
Easy to use, missing data handling occurs behind the scenes	MLM software usually discards observations with missing predictors

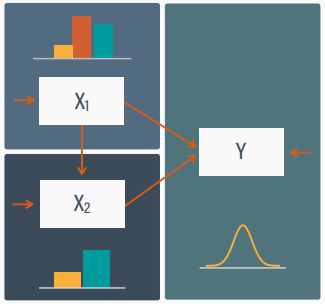
BAYESIAN ESTIMATION

THINGS BAYES ESTIMATION IS GOOD AT

- Direct estimation for complex models with missing data
- Mixed metrics (normal, ordinal, nominal, skewed, count, latent)
- Nonlinear effects (interactions, curvilinear effects)
- Multilevel data (random coefficients, interactions)
- Latent variable modeling (interactions)

FACTORED REGRESSION SPECIFICATIONS

- Factored regression specifications invoke a unique distribution for each variable
- The analysis consists of a collection of univariate regression models
- Each model can include terms that are at odds with multivariate normality



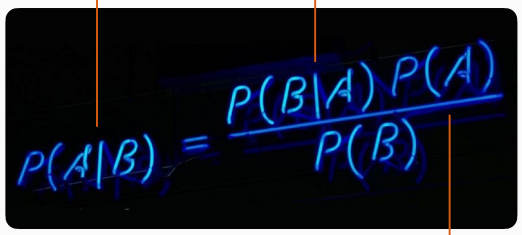
FREQUENTIST VS. BAYESIAN PARADIGMS

Frequentist	Bayesian
The parameter is a fixed quantity, estimates vary across different samples	Parameters are random variables with a distribution of plausible realizations
Statements about probability, precision, and confidence refer to estimates	Statements about probability, precision, and intervals refer to the parameter
Probability = long run frequency of outcomes across many samples	Probability = our degree of certainty about a parameter after analyzing data

BAYES' THEOREM

Posterior = parameters (A) given the data (B)

Likelihood = data (B) given the parameters (A)



Prior = a priori belief about parameters (A)

MCMC ESTIMATION

Estimate regression models

Impute missing values

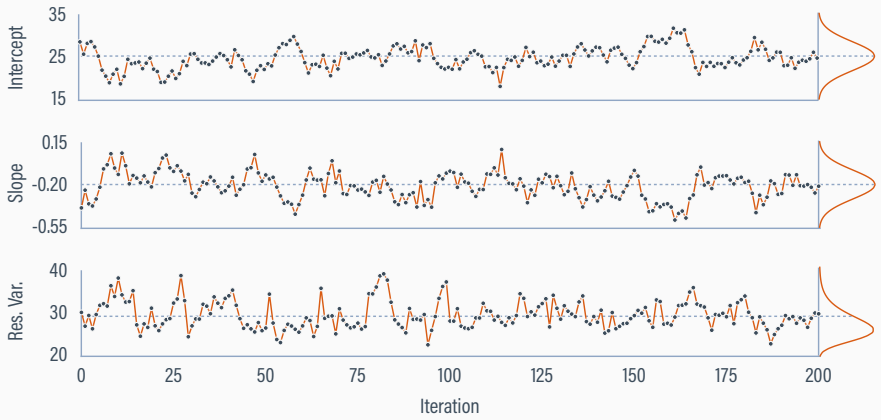
Do for $t = 1$ to 10,000 iterations

- Estimate model parameters, conditional on the filled-in data
- Impute missing values, conditional on the model parameters

Repeat

Summarize model parameters

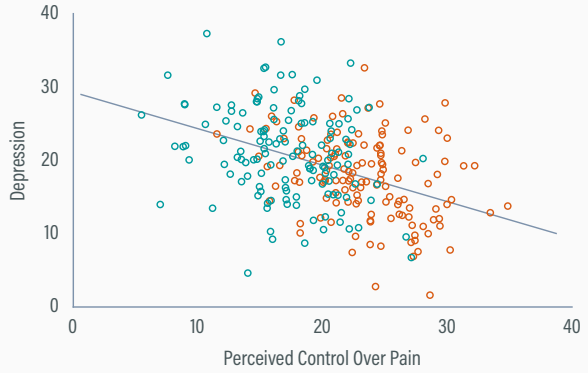
PARAMETERS FROM 200 MCMC CYCLES



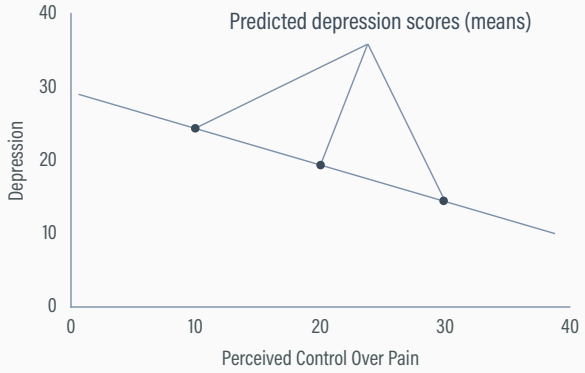
MISSING DATA IMPUTATION

- Missing scores are imputed by drawing replacement scores at random from a distribution of plausible values
- The model parameters combine to define the center and spread of the missing data imputations
- Each iteration yields unique model parameters and unique imputations based on those parameters

REGRESSION FROM FILLED-IN DATA

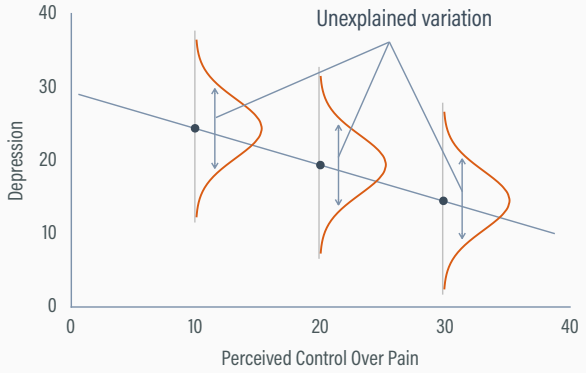


PREDICTED VALUES



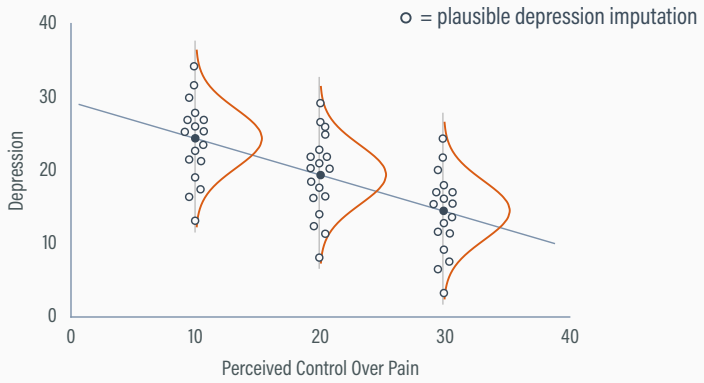
47

RESIDUAL VARIATION



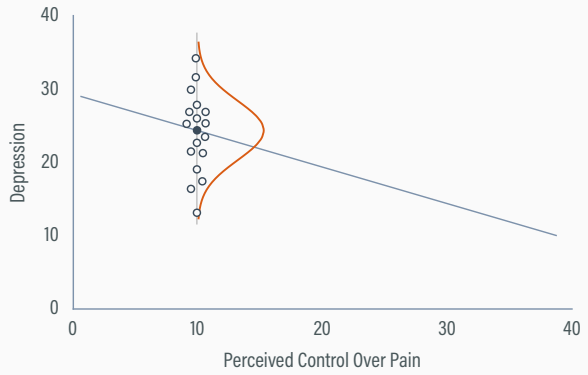
48

DISTRIBUTIONS OF IMPUTATIONS



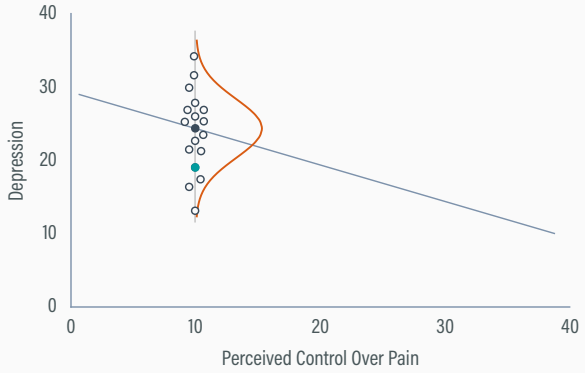
49

IMPUTATION FOR LOW PERCEIVED CONTROL



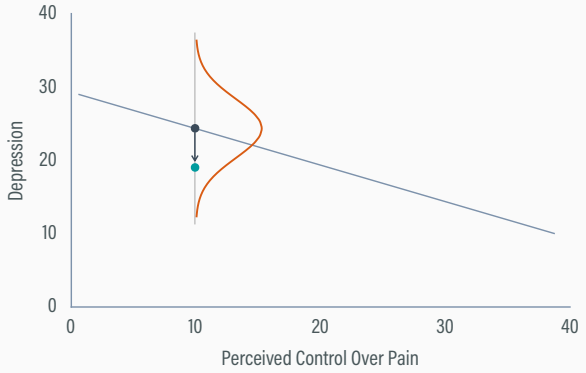
50

DRAW AN IMPUTATION AT RANDOM



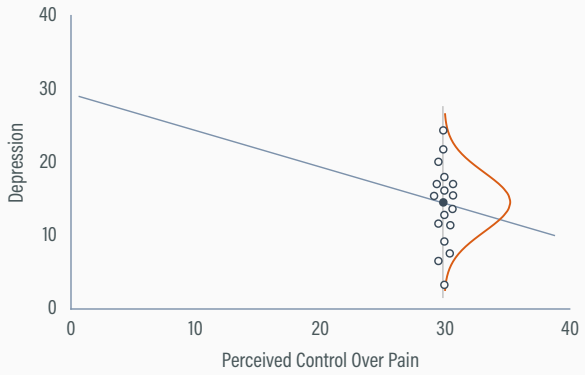
51

IMPUTATION = PREDICTION + NOISE



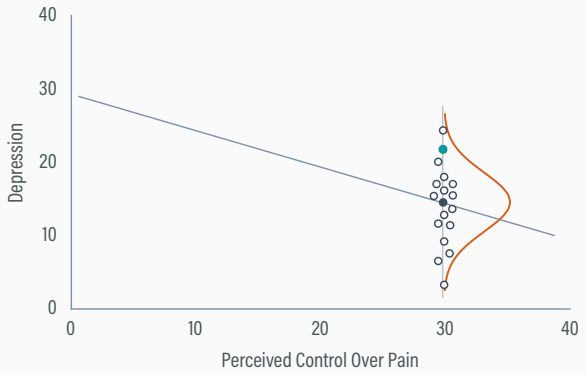
52

IMPUTATION FOR HIGH PERCEIVED CONTROL



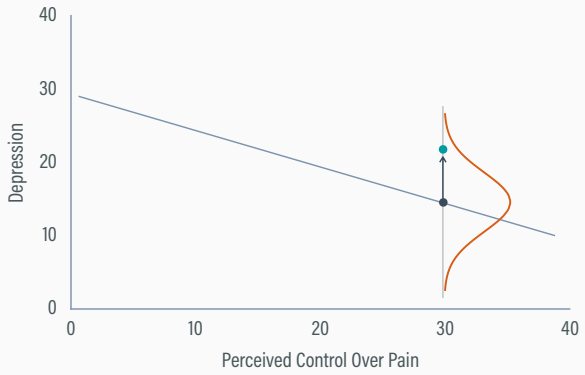
53

DRAW AN IMPUTATION AT RANDOM



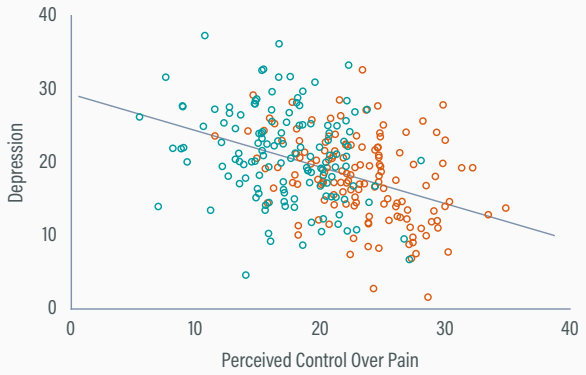
54

DRAW AN IMPUTATION AT RANDOM



55

FILLED-IN DATA AT ITERATION T



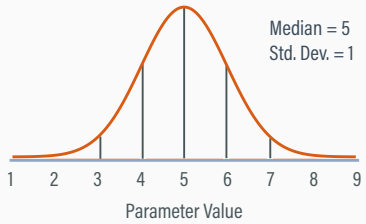
56

PARAMETER (POSTERIOR) DISTRIBUTIONS

- Bayesian estimation yields a distribution of parameters—a posterior—that averages over thousands of filled-in data sets
- The posterior describes a distribution of plausible parameter values that could have produced our particular data
- Instead of estimates varying around a fixed parameter (frequentist), parameter values vary around a fixed data set

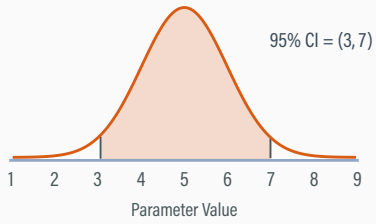
POSTERIOR MEDIAN AND STD. DEV.

- The posterior median and standard deviation quantify the most likely parameter value and uncertainty
- Analogous to a point estimate and standard error, sans repeated sampling



95% CREDIBLE INTERVALS

- The 95% credible interval gives limits spanning 95% of the parameter's range
- Akin to a confidence interval, but references a range of highly plausible parameter values



BAYESIAN ESTIMATION PROS AND CONS

Pros

- Direct estimation competitor to maximum likelihood, but more flexible
- Suited for interactions, non-linear terms, and random coefficients (MLMs)
- Good for mixed metrics (normal, binary, ordinal, multicategorical, count, skewed)

Cons

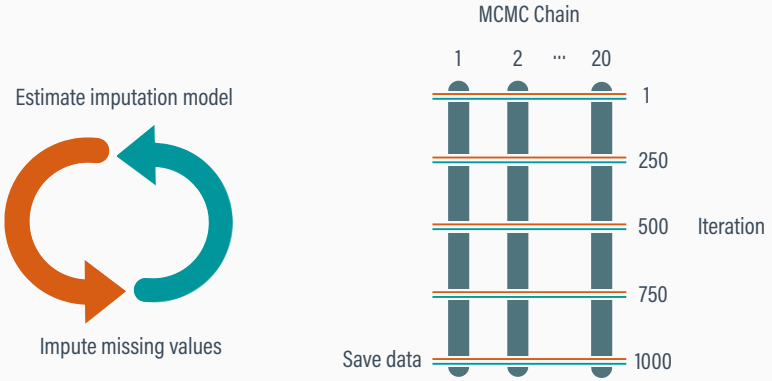
- Fewer simple software options (Blimp), most are difficult to use (JAGS)
- MCMC is not fully autonomous, requires some input and oversight
- Literature on factored regression specifications is less mature

MULTIPLE IMPUTATION

MULTIPLE IMPUTATION

- Bayesian estimation creates a filled-in data set at every iteration, and estimates average over thousands of imputations
- Multiple imputation saves a small number of data sets (e.g., 20 is common) for reanalysis using frequentist methods
- MCMC is co-opted for the purpose of creating imputations, but the Bayesian parameter estimates are not of interest

MCMC ESTIMATION



STEP 1: SAVE IMPUTED DATA SETS

Original 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

Imputed data set 1

Y	X_1	X_2
4	4	3
3	3.2	5
7	1	6
5.3	1	6
5	9	3
3	8.7	10.1
1	6	7
9	4	9
2	6.5	6

Imputed data set 2

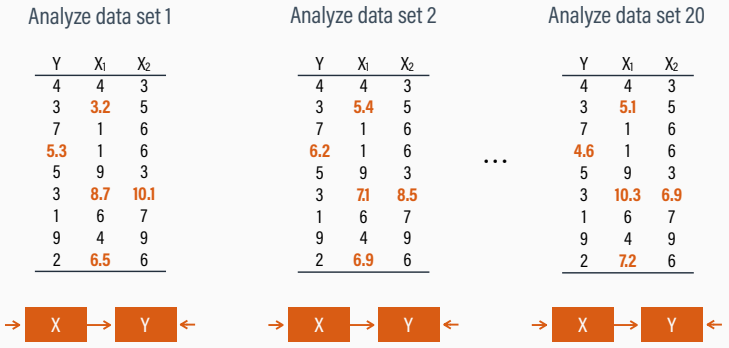
Y	X_1	X_2
4	4	3
3	5.4	5
7	1	6
6.2	1	6
5	9	3
3	7.1	8.5
1	6	7
9	4	9
2	6.9	6

...

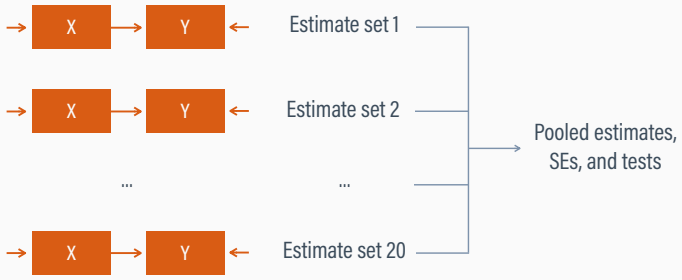
Imputed data set 20

Y	X_1	X_2
4	4	3
3	5.1	5
7	1	6
4.6	1	6
5	9	3
3	10.3	6.9
1	6	7
9	4	9
2	7.2	6

STEP 2: ANALYZE EACH DATA SET



STEP 3: POOL RESULTS

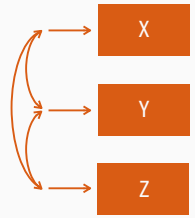


AGNOSTIC VS. MODEL-BASED IMPUTATION

- Step 1 (imputation) uses MCMC to fit a model, the parameters of which define distributions of imputations
- Step 2 (analysis) fits the focal models to the filled-in data
- Agnostic imputation deploys an imputation model that differs from the analysis model, whereas model-based imputation deploys the same model in both steps

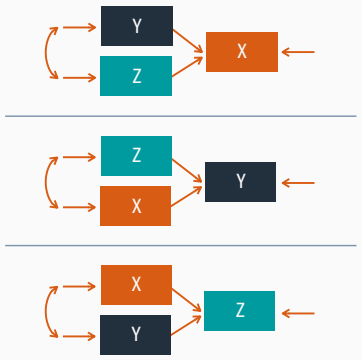
JOINT MODEL IMPUTATION

- Joint imputation invokes a multivariate distribution for the incomplete variables
- Usually a multivariate normal model with a mean vector and covariance matrix as parameters



FULLY CONDITIONAL SPECIFICATION

- FCS uses regression models to fill in data
- Each MCMC cycle uses a round-robin scheme with each variable predicted by others
- Each regression model can invoke a different metric and distribution



AGNOSTIC IMPUTATION PROS AND CONS

Pros

- Widely available in statistical software (SPSS, SAS, Stata, Mplus, R)
- Accommodates mixed metrics (normal, binary, ordinal, multicategorical)
- Can generate imputations for several purposes or analyses

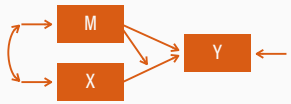
Cons

- Biased with interactions, non-linear terms, and random slope MLMs
- Capabilities vary dramatically across software packages
- Algorithms for MLMs are limited and restricted to random intercepts

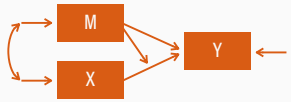
MODEL-BASED IMPUTATION

- The step 1 imputation model exactly matches the step 2 analysis model
- Imputations are tailored to one analysis, cannot be used for other purposes

Imputation model



Analysis model



MODEL-BASED IMPUTATION PROS AND CONS

Pros

- Suited for interactions, non-linear terms, and random coefficients (MLMs)
- Accommodates mixed metrics (normal, binary, ordinal, multicategorical)
- Imputation and analysis models cannot conflict or contradict each other

Cons

- Fewer simple software options (Blimp), some are difficult to use (JAGS)
- Each analysis requires a unique set of tailored imputations
- Literature on factored regression specifications is less mature

MISSING DATA SOFTWARE

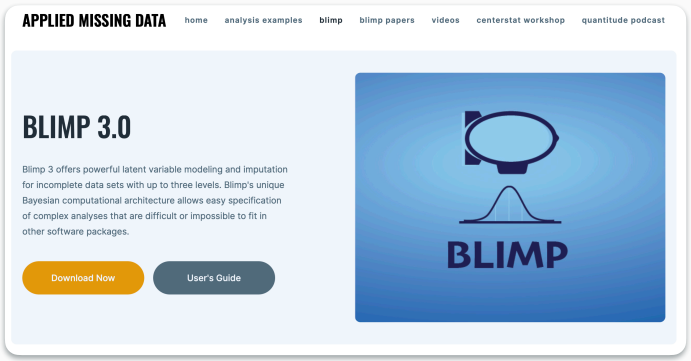
SOFTWARE RECOMMENDATIONS

Method	Program	Specification	Features
Maximum likelihood	Mplus	MVN limited FRS	binary & ordinal variables / robust corrections / random intercept MLMs / latent variable interactions / MI analyses
	lavaan	MVN	normal variables only / robust corrections / random intercept MLMs / MI analyses
Bayesian estimation	Blimp	FRS	user-friendly / binary, ordinal, multicategorical, skewed, count, latent / any MLM / any latent or manifest interaction
Multiple imputation	Blimp	model-based FRS agnostic FCS	same features as Bayesian (model-based) / FCS with binary, ordinal, multicategorical, latent response / MLMs
Imputation analysis	Mplus mitml (R)	NA	comprehensive multiple imputation analysis and pooling suites with test statistics

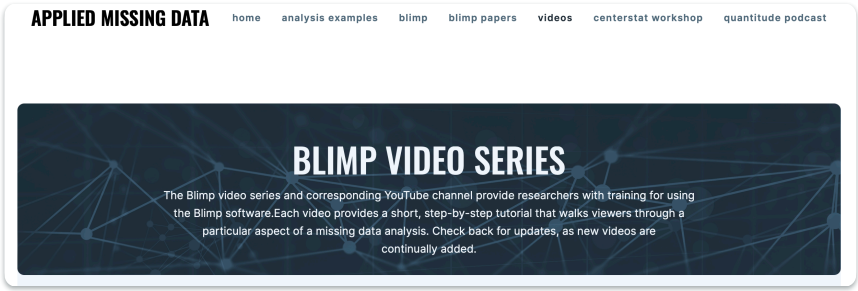
Note. MVN = multivariate normal, FRS = factored regression specification, FCS = fully conditional specification

WWW.APPLIEDMISSINGDATA.COM/BLIMP

Institute of Education Sciences



WWW.APPLIEDMISSINGDATA.COM/VIDEOS



ANALYSIS EXAMPLE

CHRONIC PAIN ANALYSIS

$$DEPRESS = \beta_0 + \beta_1(INTERFERE) + \beta_2(CONTROL) + \beta_3(PAIN) + \beta_4(AGE) + \beta_5(TXGRP) + \varepsilon$$

Variable	Definition	Missing %	Scale
DEPRESS	Depression composite	13.5	Numeric
INTERFERE	Pain interference with life composite	10.6	Numeric
CONTROL	Perceived control over pain composite	0	Numeric
PAIN	Severe pain dummy code	7.3	0 = No/little pain, 1 = Severe pain
AGE	Age in years	0	Numeric
TXGRP	Treatment assignment dummy code	0	0 = Waitlist control, 1 = Treatment

BIG THREE COMPARISON

The Big Three are numerically equivalent!!!

	Bayes		FIML		Multiple Imputation	
Parameter	Est.	SD	Est.	SE	Est.	SE
Intercept	18.07	2.68	18.14	2.63	17.92	2.57
Pain interference slope	0.13	0.05	0.13	0.05	0.14	0.05
Perceived control slope	−0.22	0.08	−0.22	0.08	−0.23	0.08
Severe pain slope	1.39	0.90	1.38	0.89	1.30	0.90
Age slope	−0.09	0.03	−0.09	0.03	−0.08	0.03
Treatment group slope	1.99	0.75	1.99	0.73	2.03	0.71
R^2	.20	.04	.20	.05	.20	.04

For more information go to

WWW.APPLIEDMISSINGDATA.COM