Module 9

Introduction to Missing Data Methods

Katie Wilson, PhD
Department of Biostatistics, University of Washington

SISCER 2024
July 18-19, 2024

A Bit About Me

Katie Wilson

Dept. of Biostatistics
Univ. of Washington

Thank you to Dr. Anna Plantinga (Williams) who co-created this module.

Resources

- **Module website:** slides, R code, schedule
- **Slack:** ask questions, interact with other module participants, access recordings
- **Office hours:** 1-2pm PT on Friday, July 19
- I'll recommend **textbooks and articles** for further reading throughout and at the end of the module

Goals of this module

- Overview of missing data challenges
- Study design considerations
- Definitions for key terms
- **Main focus: introduce statistical approaches for analyzing data with missingness**
- How to use R to run these analyses

Today, July 17, 8:30-12:00

8:30-9:30 Introduction to missing data + some key terminology
9:30-9:45 Break
9:45-10:45 Conventional methods
10:45-11:00 Break
11:00-12:00 Maximum likelihood

Tuesday, July 18, 8:30-12:00

8:30-9:30 Multiple imputation
9:30-9:45 Break
9:45-10:45 Inverse probability weighting
10:45-11:00 Break
11:00-11:30 Sensitivity analyses
11:30-12:00 Wrap-up

A Bit About You

Please type in the chat:

- Briefly introduce yourself - what's your name, and what state or country are you currently located in?
- What's your primary role in most of the studies in which you participate?

Overview

What Do We Mean by "Missing" Data??

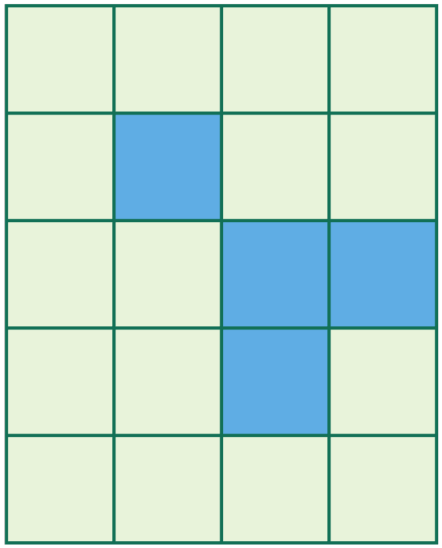
Missing data are unobserved values that would be meaningful for analysis if observed; in other words, a missing value hides a meaningful value

(Little & Rubin)

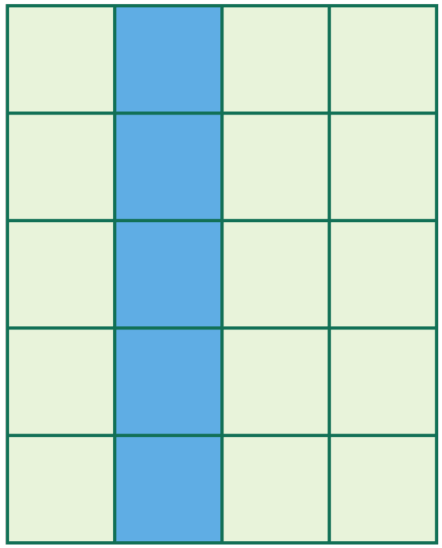
- **Example 1**: participant in a survey does not report income
- **Example 2**: heart rate monitor is not working
- **Example 3**: dropout (where participants stop showing up for visits after a certain point) in a longitudinal study
- **Example 4**(??): participant in a political poll not able to express preference for a particular candidate
  - Suppose question is "who will you vote for?"
  - People who are "missing" may be (1) not going to vote or (2) refusing to provide a response

What Do We Mean by "Missing" Data??

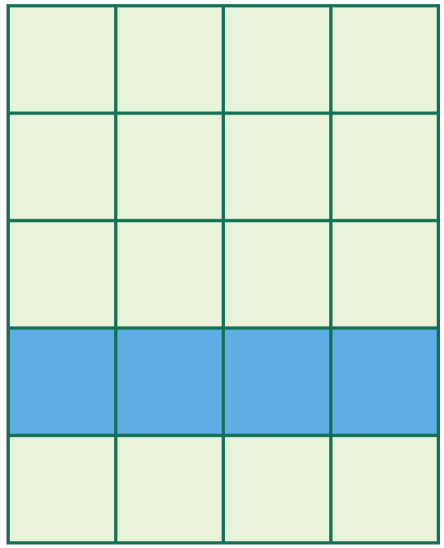
Item non-response



Latent variable



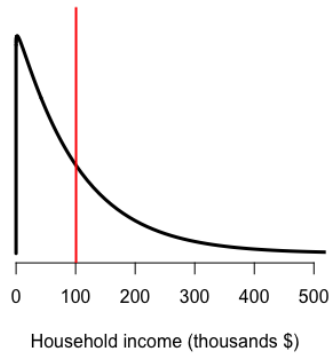
Unit non-response

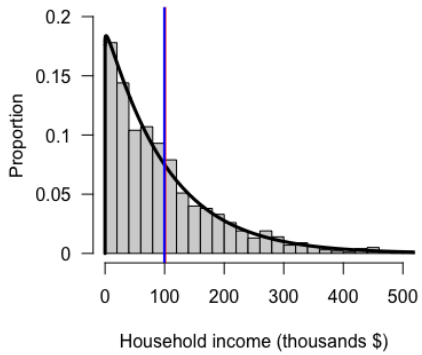


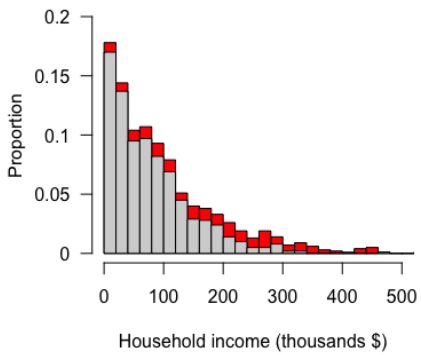
We will be focusing on *item* non-response (data missing for some, not all cases); however, methods can be adapted for latent variable and *unit* non-response

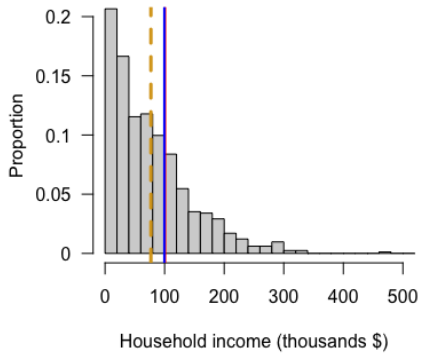
Why Is Missing Data a Problem?

Income example:









- L: probability density of outcome of interest
- ML: histogram of ideal data from all 1000 people in the sample
- MR: histogram showing what fraction is missing
- R: histogram of observed data from 823 people who responded

Why Is Missing Data a Problem?

GOAL: make inference about some aspect of an underlying population of interest using data that would have been observed had there been no missingness

- Standard statistical methods assume relevant variables are all observed
- Loss of precision
- Potential for bias
  - Can we view the 823 responses as a random sample from the population?
  - Or are those with higher incomes less likely to respond?

⋆ **Major challenge**: the process by which observations become missing is not usually known and any assumptions are not possible to verify from the data available

So... What Can We Do About It?

Start at **study design**! By planning, we can reduce bias, increase efficiency, and improve whole study process (clarity of scientific question, study design, data collection, analysis)

1. **Minimize amount of missing data**
   - Clearly define scientific question, focus on obtaining the data necessary to address
   - For studies with follow-up, consider reducing follow-up time, increase schedule flexibility
   - Improve communication with study participants (e.g., multiple modes of contact)
2. **Reduce impact of missing data**
   - Collect reasons for missing data
   - Collect **auxiliary** data that is predictive of the variable with missingness
   - Collect secondary outcomes that are easier to obtain

Cross-Sectional Designs and Longitudinal Designs

Cross-Sectional:

ID	X_1	X_2	Y
1	4.67	0	5.16
2	3.35		4.73
3	1.63	1	
4	3.96	0	4.96
5		0	-0.6
6	1.21	1	4.26

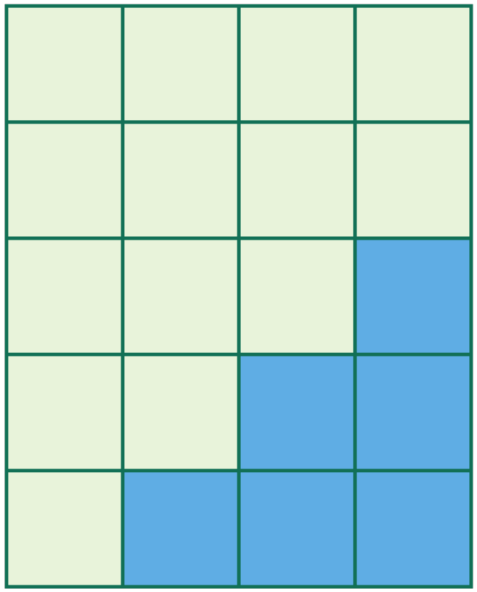
Longitudinal:

ID	X_1	X_2	Y_1	Y_2	Y_3
1	5.20	1	1.23	2.31	1.87
2	4.62	0	0.73	0.81	0.82
3	6.45	1	0.52		
4	2.88	1	1.02	1.45	1.68
5	3.56	0	0.92	0.86	
6	4.21	0	1.45	1.21	

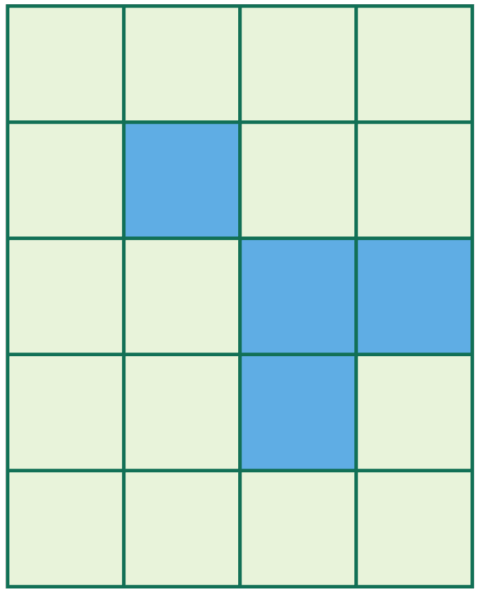
In this module, our focus will be on cross-sectional studies, but will point out how to use missing data approaches in the context of longitudinal designs

Definitions I: Patterns of Missingness

Monotone



Non-monotone

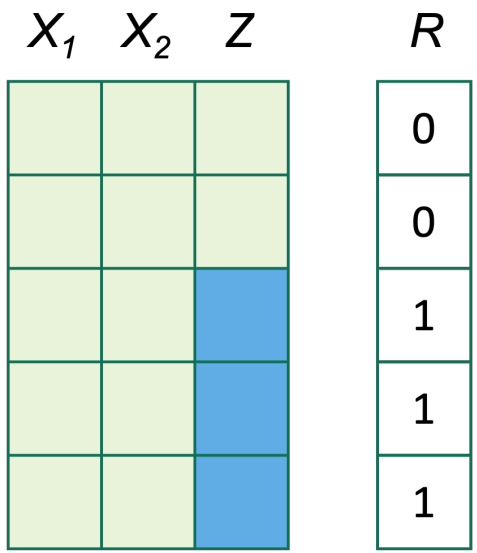


- Monotone patterns can be easier to work with in analyses
- Monotone patterns arise in longitudinal studies with *dropout*

Notation

Helpful notation we will use:

- X, Z, Y, U, R all represent variables
  - Usually, Xs will be thought of as predictors/covariates
  - Usually, Y will be thought of as the outcome/response
  - Z represents a variable that has missingness (could be an outcome or predictor)
  - U will denote unmeasured variables
  - R is an indicator for missingness
    - $R = 1$ if Z is missing



- ϕ: parameters that govern the missing data process
- θ: parameters that are of scientific interest (e.g., regression coefficients from the "analysis" model)

Definitions II: Missing Data Mechanisms; Rubin (1976)

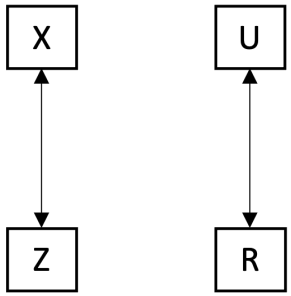
In order to obtain valid inference from incomplete data need to consider the mechanism (probability model) producing the missing data.

What are assumptions we could make about the missing data?
Informally,

1. **Missing Completely at Random (MCAR)**: missingness does not depend on other observed variables nor the value of the missing variable itself
2. **Missing at Random (MAR)**: missingness depends only on other observed variables but not the missing variable itself
3. **Missing Not at Random (MNAR)**: missingness depends on missing values

Missing Data Mechanisms: MCAR

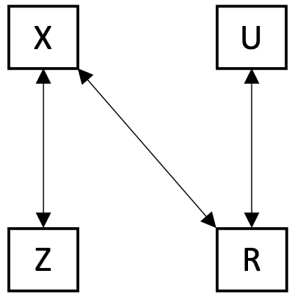
$$\Pr(R = 1|Z, \mathbf{X}, \phi) = \Pr(R = 1|\phi)$$



- The probability that Z is missing does not depend on observed variables $\mathbf{X}$ or missing value of Z
- Safe to assume MCAR when *missing by design*: e.g., by design only obtain measurements for Z on random subsample of the data
- Example: study of physical activity involving activity trackers and the activity tracker does not record one day
- While it is possible to test whether missingness depends on $\mathbf{X}$, it is not possible to test whether missingness depends on Z

Missing Data Mechanisms: MAR

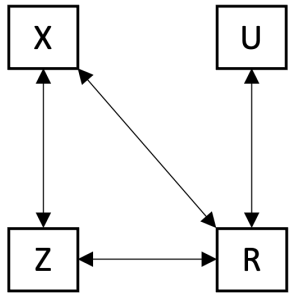
$$\Pr(R = 1|Z, \mathbf{X}, \phi) = \Pr(R = 1|\mathbf{X}, \phi)$$



- The probability that Z is missing may depend on $\mathbf{X}$, but does not depend on Z (after adjusting for $\mathbf{X}$)
- Example: missing income depends on occupation, but within occupation groups probability of missing income does not depend on income
- Not testable
  - Make more plausible by including more variables in $\mathbf{X}$

Missing Data Mechanisms: MNAR

$$\Pr(R = 1|Z, \mathbf{X}, \phi)$$

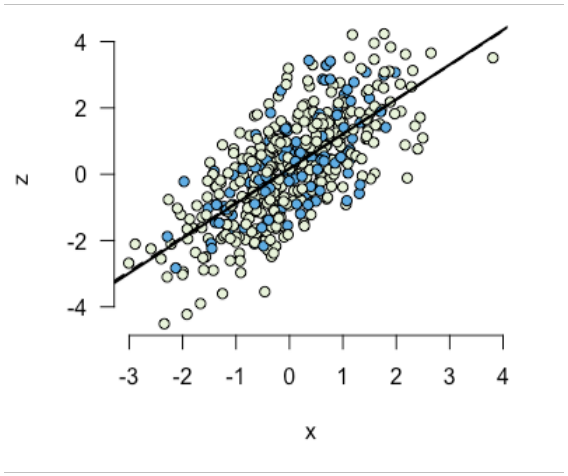


- The probability that Z is missing depends on value of Z
- Example: longitudinal study where the outcome is a measure of quality of life and participants do not show up when their quality of life is compromised
- Note: this association can occur for two reasons
  1. The probability of missing data is directly related to Z
  2. An unmeasured variable induces correlation between Z and missingness

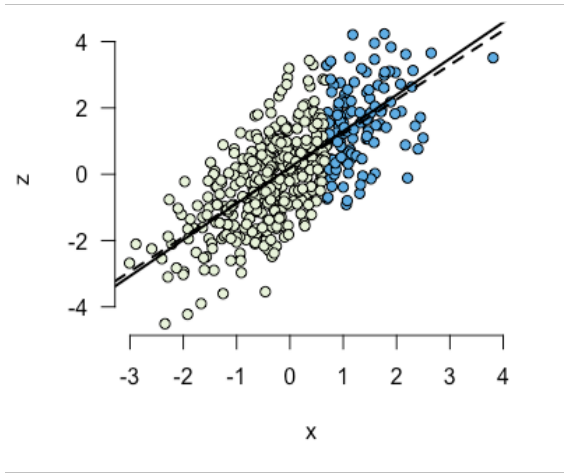
Example

- X is fully observed
- Z is observed for the lighter green points, missing for the darker blue points

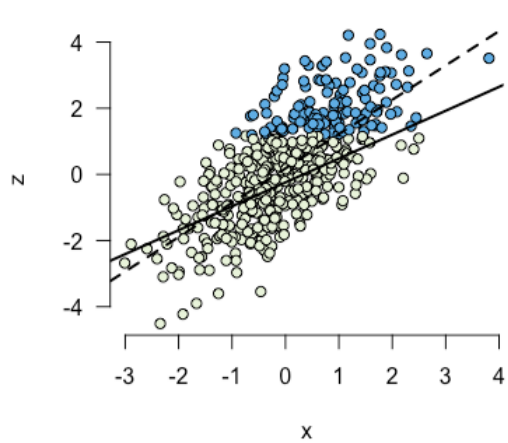
MCAR



MAR



MNAR



Definitions III: Ignorable and Non-ignorable

- Ideally, we would like to ignore the parameters ϕ that govern the missing data process
- When can we get valid inference for the parameters of scientific interest, θ, without knowing ϕ?
  - If missing data mechanism is **ignorable** then possible to get valid inference without directly modeling the missing data mechanism
  - **Note**: this does not necessarily imply that we can just remove anyone with missing data from the analysis

- **Ignorable**: data MAR and a technical condition that is unlikely to be violated in practice is met (that is, parameters in analysis model are separate from those that govern the missingness mechanism)
- **Non-ignorable**: MNAR

Plausibility of MAR

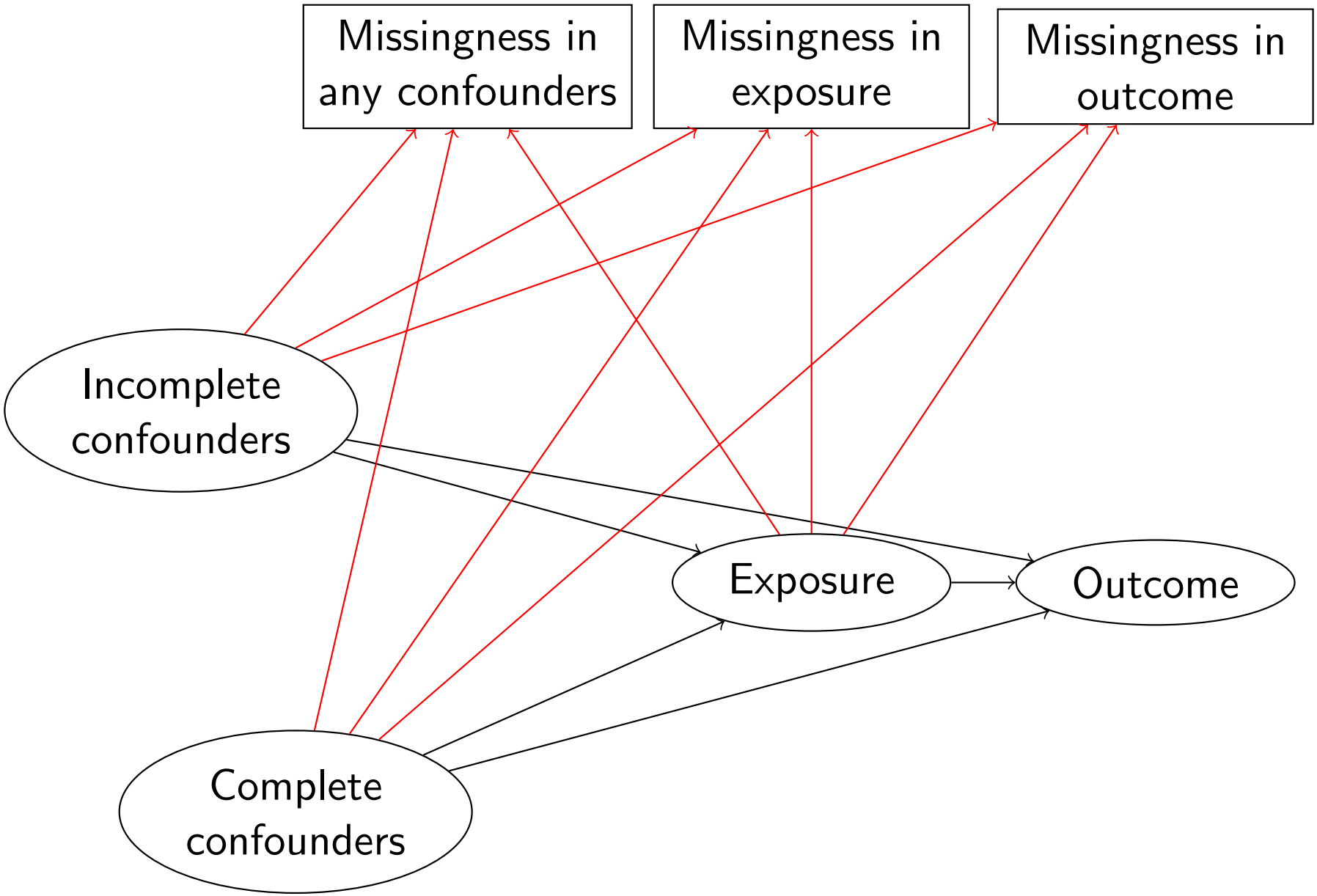
In this module, we'll usually be assuming MAR.

- Sometimes, MAR known to hold, such as in cases with *planned missingness*
  - Viewing randomized experiment as a missing data problem (unobserved response to other treatment assignment is MCAR)
  - Having only some participants fill out some questionnaire items
  - Participants included in a follow-up visit if an earlier measurement exceeded a certain threshold
- Otherwise, MAR is an assumption
  - Should generally anticipate that this may not hold
  - We'll discuss sensitivity analyses that can be done later!

Using DAGs to Assess Missingness

- When multiple variables have missingness, MAR is challenging to assess
- A **directed acyclic graph (DAG)** with additional nodes for missingness in variables can be very helpful
  - Use to determine whether the parameter of interest is possible to estimate from the available data ("recoverable")
  - Help decide what approach to handling missing data is appropriate
- References and resources:
  - Lee, K. J. et al. (2023). Assumptions and analysis planning in studies with missing data in multiple variables: moving beyond the MCAR/MAR/MNAR classification. *International Journal of Epidemiology*. `https://doi.org/10.1093/ije/dyad008`
  - Moreno-Betancur, et al. (2018). Canonical causal diagrams to guide the treatment of missing data in epidemiologic studies. *American Journal of Epidemiology*. `https://doi.org/10.1093/aje/kwy173`

Example DAG

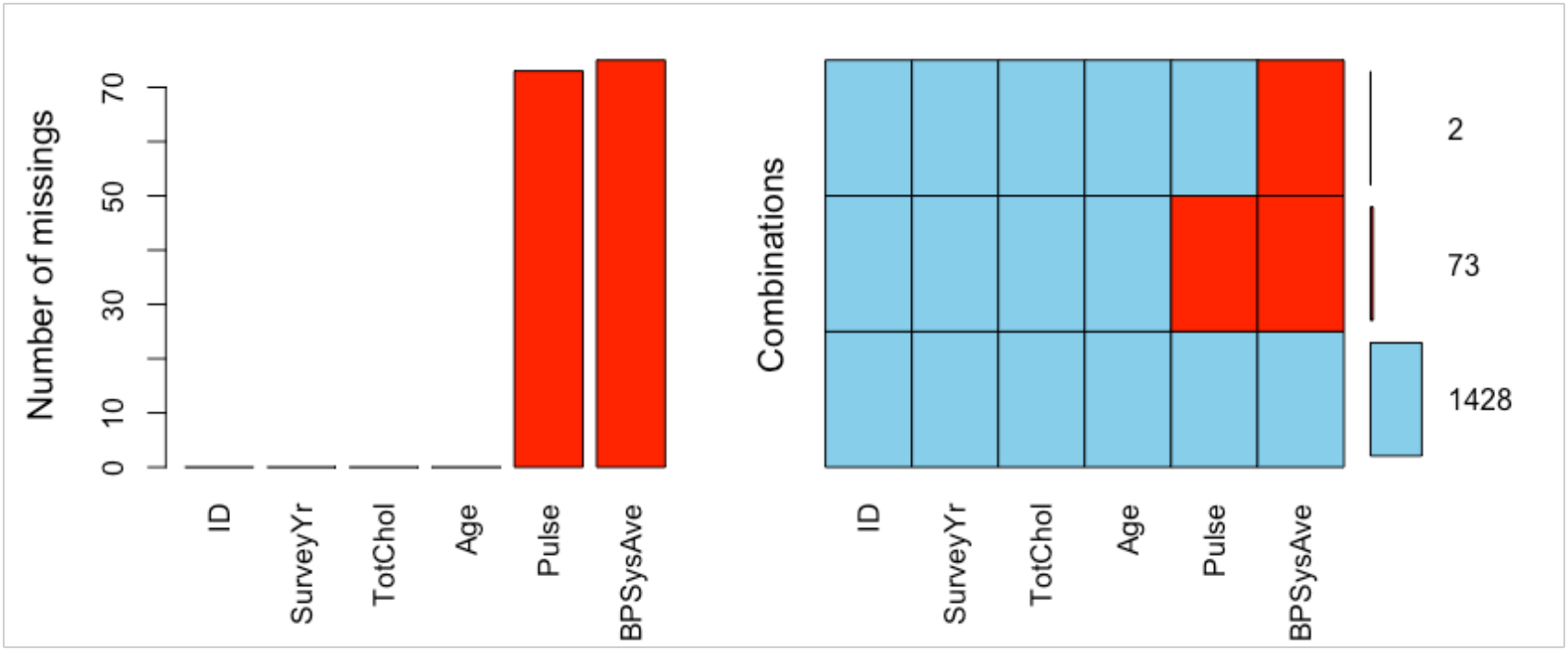

Missingness in any confounders
Missingness in exposure
Missingness in outcome
Incomplete confounders
Exposure
Outcome
Complete confounders

Case Study I: NHANES

- Survey data collected from the American National Health and Nutrition Examination surveys (NHANES)
- Dataset we will use is a resampled version of the original data and for our purposes can be treated as a simple random sample from the US population
- Restrict to $\approx$ 1,500 non-pregnant female individuals ages 20-44, surveyed 2009 - 2012
- Outcome of interest: total cholesterol
- Covariates of interest: age, pulse, and systolic blood pressure
  - Pulse and systolic blood pressure missing for 5% of participants

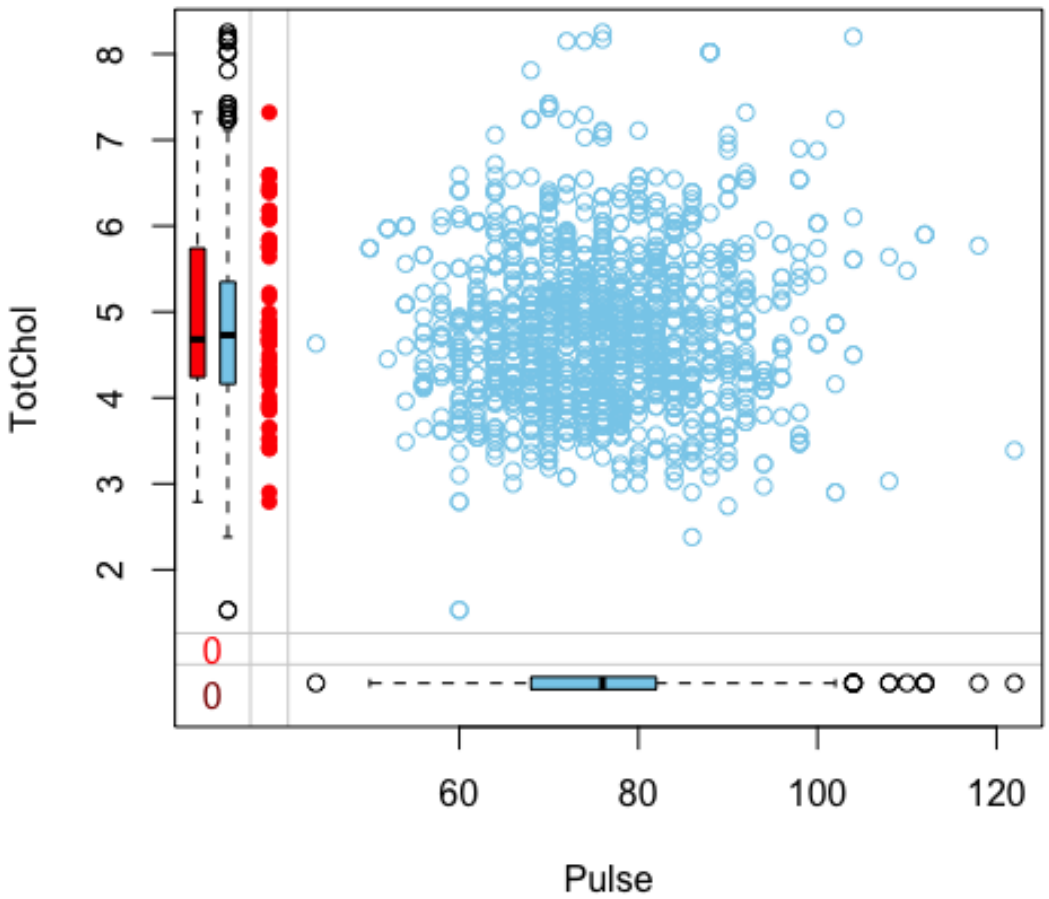
Case Study I: NHANES

The `VIM` package in `R` contains the `aggr()` function:



Case Study I: NHANES

The `VIM` package in `R` contains the `marginplot()` function:



Case Study II: NACC Uniform Data Set

THE NIA ALZHEIMER'S DISEASE RESEARCH CENTERS PROGRAM

National Alzheimer's Coordinating Center

- National Alzheimer's Coordinating Center (NACC) Uniform Data Set available here: `https://naccdata.org/`
- Longitudinal study started in 2005, participants were recruited or referred for evaluation of dementia
- In our analysis, we included the initial visit of $\approx$ 24,000 individuals who were 65+

Case Study II: NACC Uniform Data Set

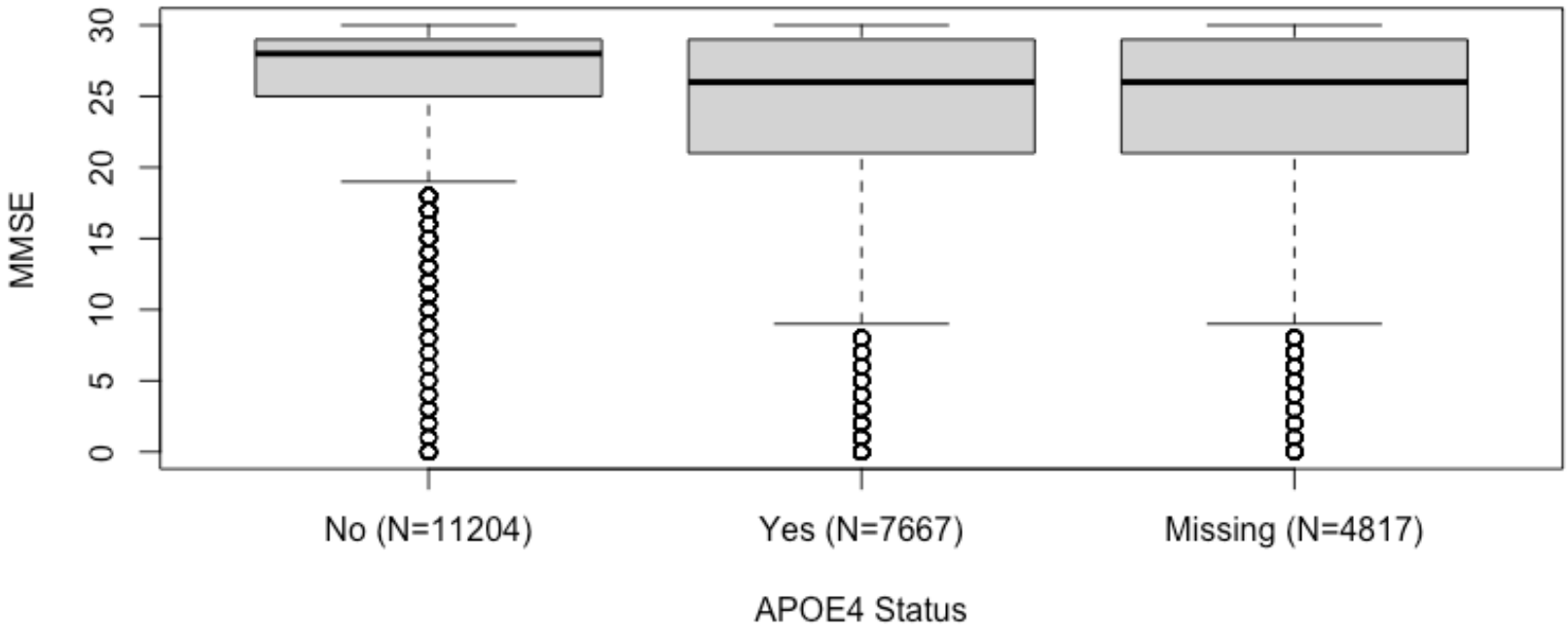
- Outcome of interest: mini-mental state examination (MMSE) score
  - The MMSE is a test used to check for cognitive impairment (lower scores indicate more severe impairment)
- Goal: investigate risk factors that impact MMSE score and whether the patient's apolipoprotein E e4 (APOE4) allele status modifies these relationships
  - APOE4 allele increases the risk for Alzheimer's; missing for $\approx 20\%$; assume missingness may depend on demographics, but not status itself
  - Risk factors: age at MMSE, sex, clinical diagnosis of AD, history of stroke, Parkinson's disease, depression
- Similar analysis to Zhou, X. H. et al (2014). Applied missing data analysis in the health sciences.

Case Study II: NACC Uniform Data Set

Table of risk factors stratified by participant's e4 allele status (number of copies):

	No ($N = 11,377$)	**At least one** ($N = 7,876$)	**Missing** ($N = 4,809$)
Sex (% Female)	57%	55%	58%
Age	76.6 (7.3)	75.3 (6.6)	77.2 (7.0)
MMSE	26.3 (5.0)	24.1 (6.4)	23.6 (6.7)
Clinical AD (% Yes)	29%	53%	48%
Stroke history (% Yes)	5%	5%	8%
Parkinson's disease (% Yes)	2%	1%	4%
Depression (% Yes)	15%	17%	22%

Case Study II: NACC Uniform Data Set



Overview of Approaches Discussed

- Naive/conventional approaches: listwise deletion, pairwise deletion, dummy variable adjustment, last observation carried forward, single imputation
- Modern approaches: maximum likelihood, multiple imputation, inverse probability weighting

As we discuss, we'll consider the following criteria in evaluating the approach's performance:

- Minimizing bias
- Maximizing use of available data
- Obtaining valid estimates of uncertainty

Overview

Conventional and Naive Methods for Missing Data

- (A selection of) conventional/naive missing data methods:
  - Listwise deletion (complete case analysis)
  - Pairwise deletion (available data analysis)
  - Dummy variable adjustment (missing-indicator method)
  - Last observation carried forward
  - Single imputation methods

- We'll outline each approach and describe when it does (and doesn't!) work

Assessing Missing Data Approaches

- What are we looking for in an approach to missing data?
- Requirements for a **valid** approach:
  - **Consistent**: Estimates are approximately unbiased in large samples
    - We will look at bias, $E(\hat{\theta} - \theta)$, in simulated datasets with $n = 500$
  - **Valid standard error estimation**: Estimate of SE should yield expected inferential properties
    - We will look at confidence interval coverage (nominally 95%)
- Among valid approaches, maximize **efficiency**
  - An estimator is more efficient if it can achieve the same standard error with fewer data points
  - Two possible measures for this:
    - The standard deviation of estimates across simulated datasets
    - The width of the 95% confidence interval

Simulation Set-Up: Full Data

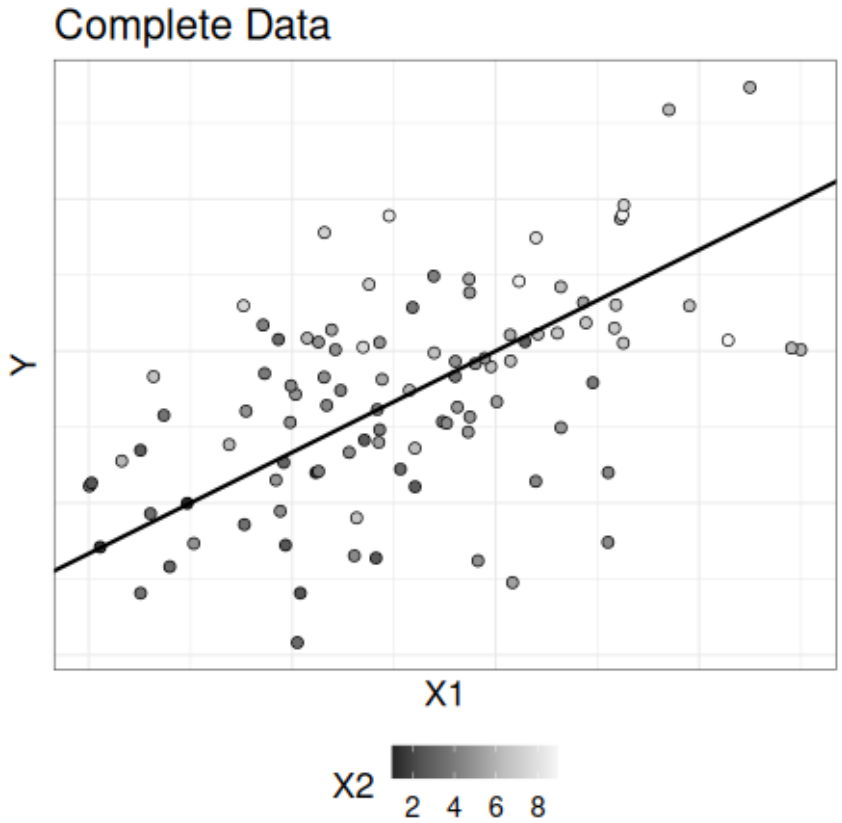
- Complete data:

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 5 \\ 5 \end{bmatrix}, 3 \cdot \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

$$Y \sim N(0.5 \times X_1 + 0.5 \times X_2, \sigma^2)$$

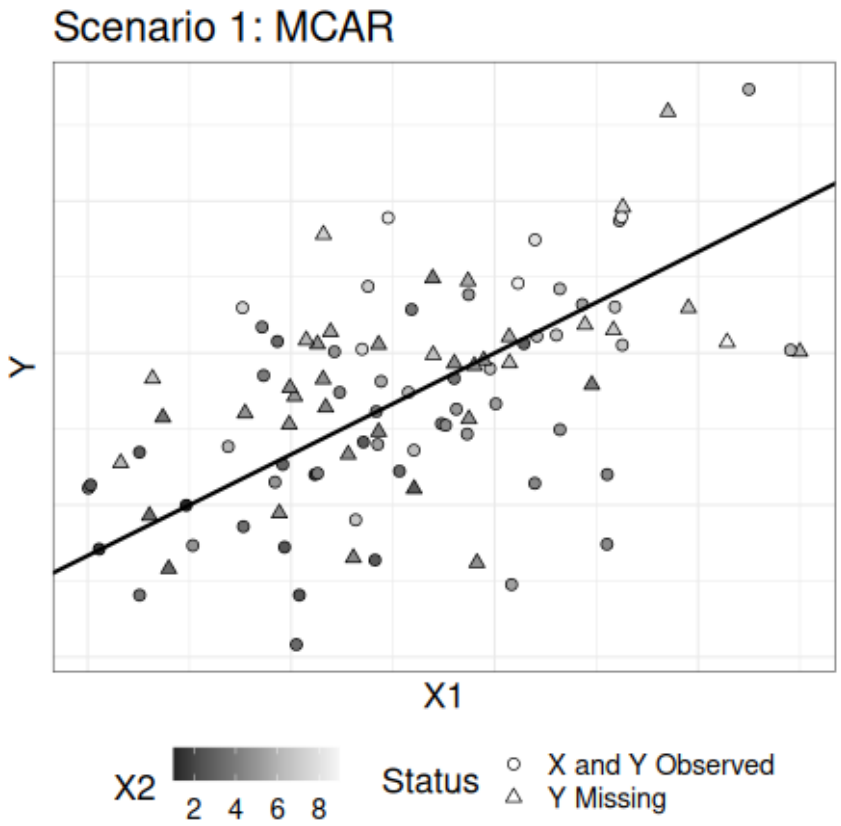
- Goal: estimate regression coefficients in

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$



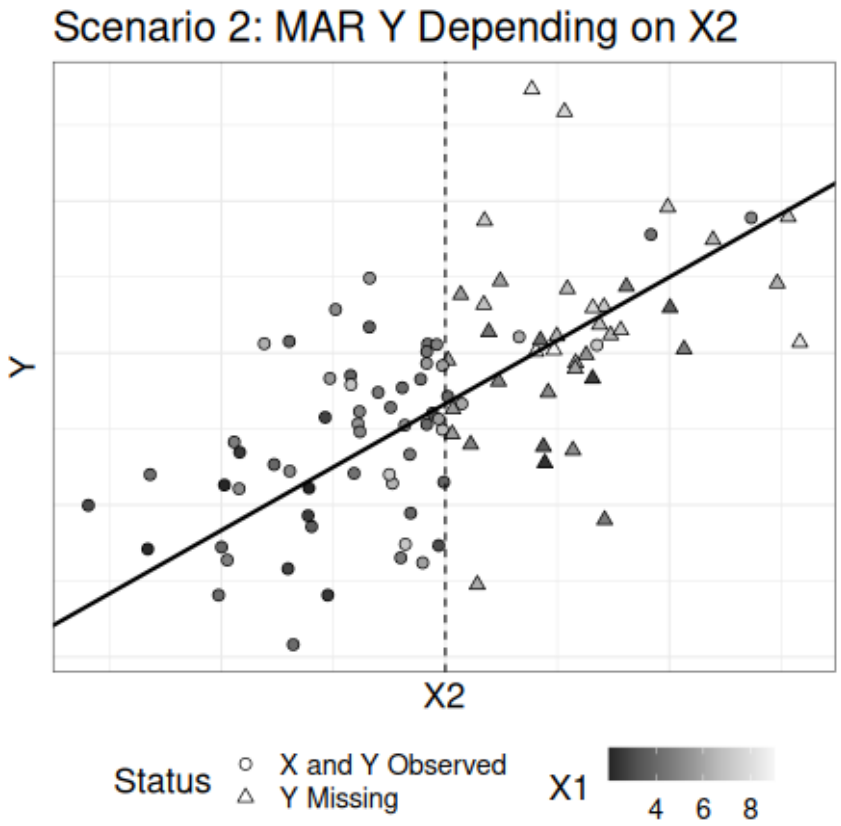
Simulation Scenario 1: MCAR

- Missingness only in Y
- Missing Completely at Random
  - Constant $P(R = 1) = 0.4$



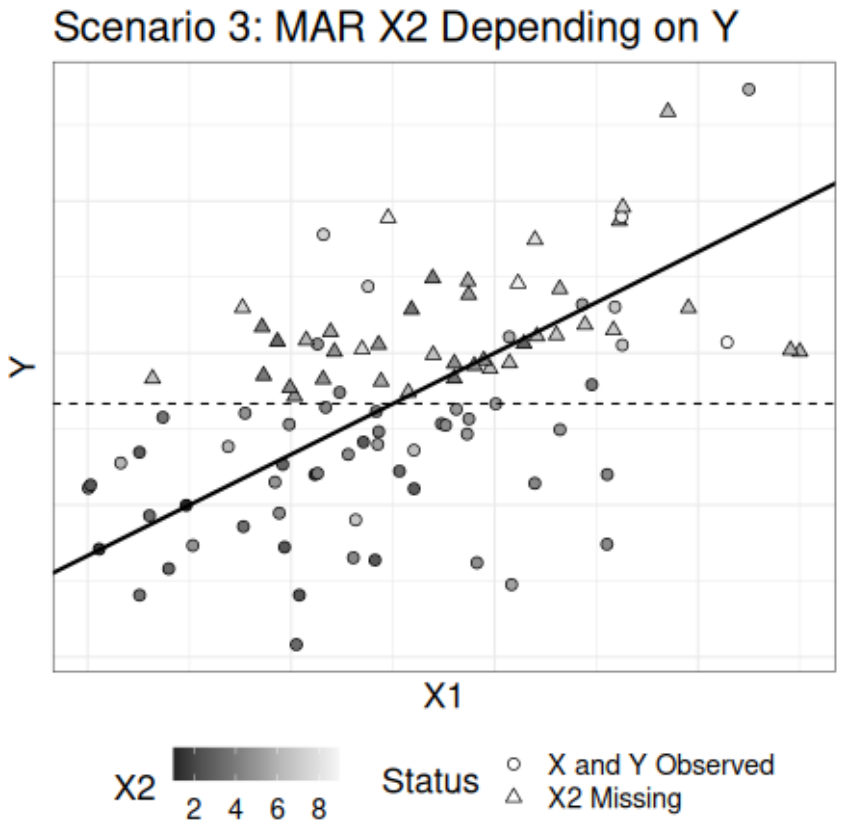
Simulation Scenario 2: Missingness Depends on X_2

- Missingness only in Y
- Missing at Random depending on value of X_2
  - $P(R = 1|X_2 \geq 5) = 0.8$
  - $P(R = 1|X_2 < 5) = 0$
- Total proportion of observations with missing data is 40%



Simulation Scenario 3: Missingness Depends on Y

- Missingness only in X_2
- Missing at Random depending on value of Y (observed!)
  - $P(R = 1|Y > 5) = 0.8$
  - $P(R = 1|Y \leq 5) = 0$
- Total proportion of observations with missing data is 40%



Conventional Approaches

- Listwise deletion (complete case analysis)
- Pairwise deletion (available data analysis)
- Dummy variable adjustment (missing-indicator method)
- Last observation carried forward
- Single imputation methods

Listwise Deletion (Complete Case Analysis)

- **Cross-sectional:** Remove subjects with missing data *on any variable* used for an analysis in this paper

ID	X_1	X_2	Y
1	4.67	0	5.16
2	3.35		4.73
3	1.63	1	
4	3.96	1	4.96
5		0	-0.6
6	1.21	1	4.26

$\rightarrow$

ID	X_1	X_2	Y
1	4.67	0	5.16
4	3.96	1	4.96
6	1.21	1	4.26

Listwise Deletion (Complete Case Analysis)

- **Longitudinal:** Remove *all time points* for subjects with missing data

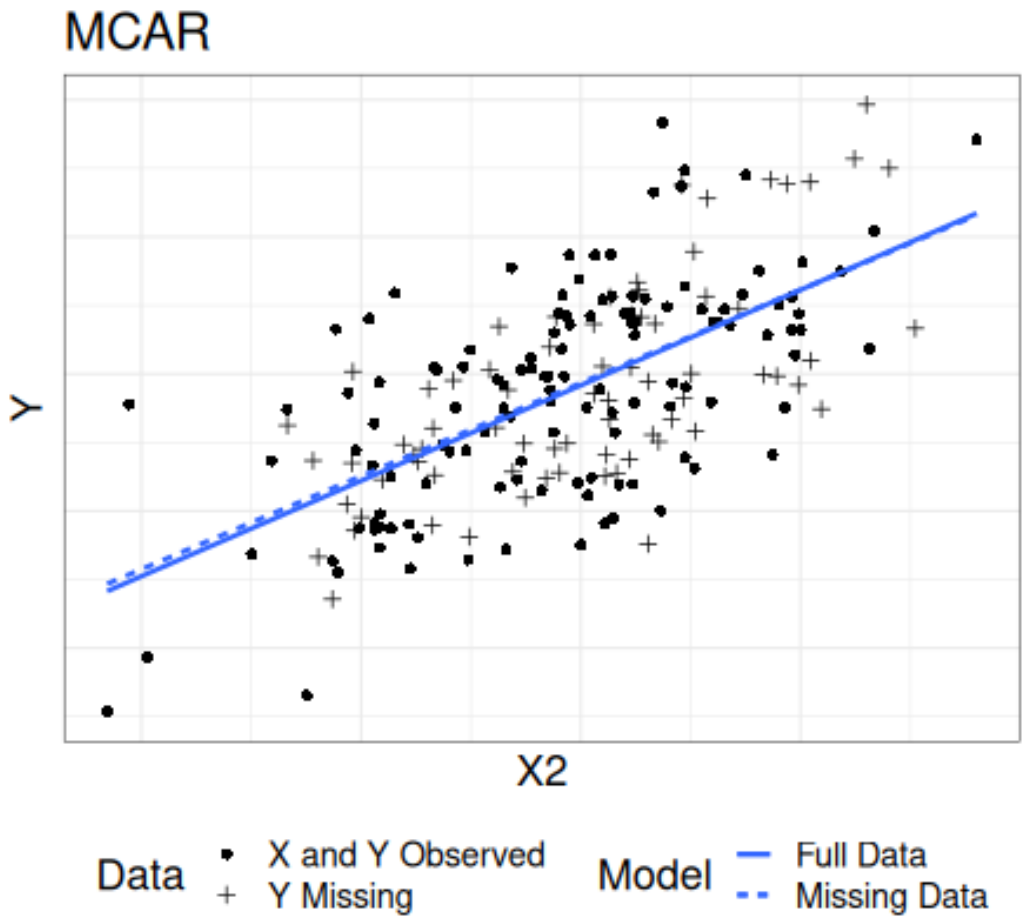
ID	Time	X_1	X_2	Y
1	1	4.25	1	8.53
1	2	3.61	1	
1	3	2.98	1	
2	1	4.93	0	5.58
2	2	4.62	0	7.21
2	3	5.84	0	11.64
3	1	4.28	0	6.42
3	2	5.63	0	10.22
3	3	4.64	0	

$\rightarrow$

ID	Time	X_1	X_2	Y
2	1	4.93	0	5.58
2	2	4.62	0	7.21
2	3	5.84	0	11.64

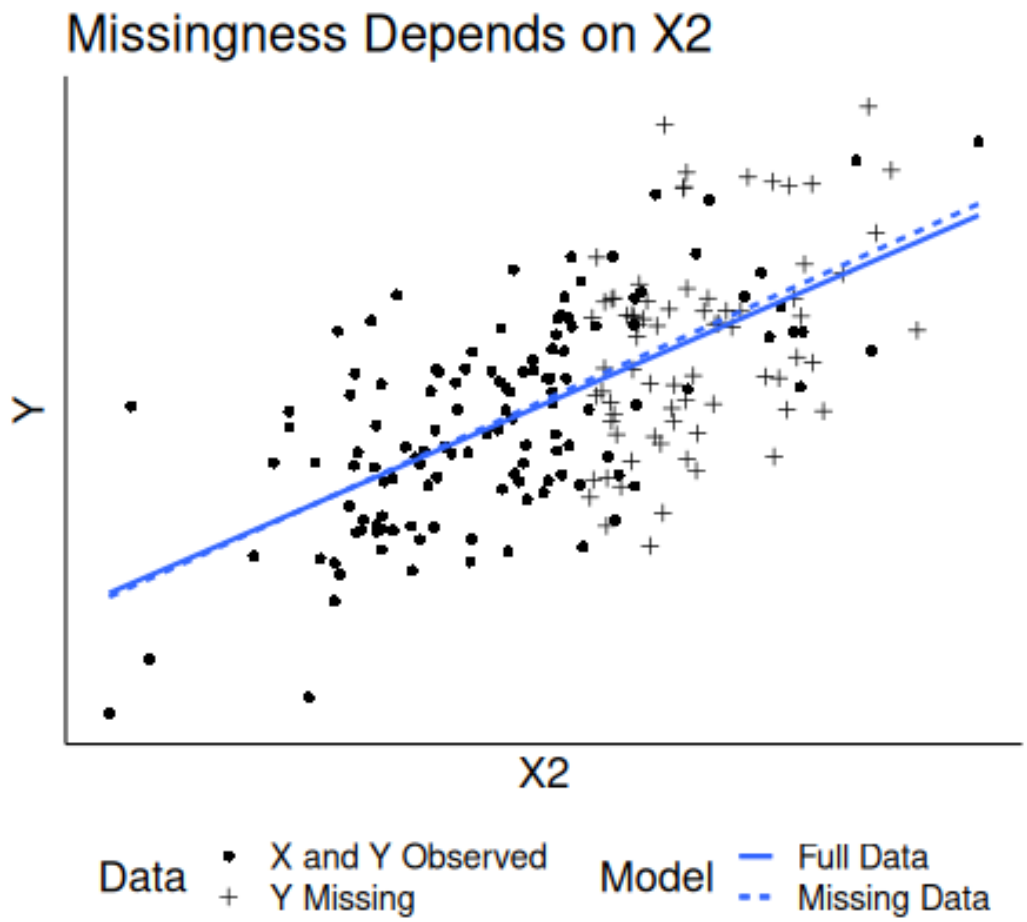
Listwise Deletion (Complete Case Analysis)

- MCAR: unbiased



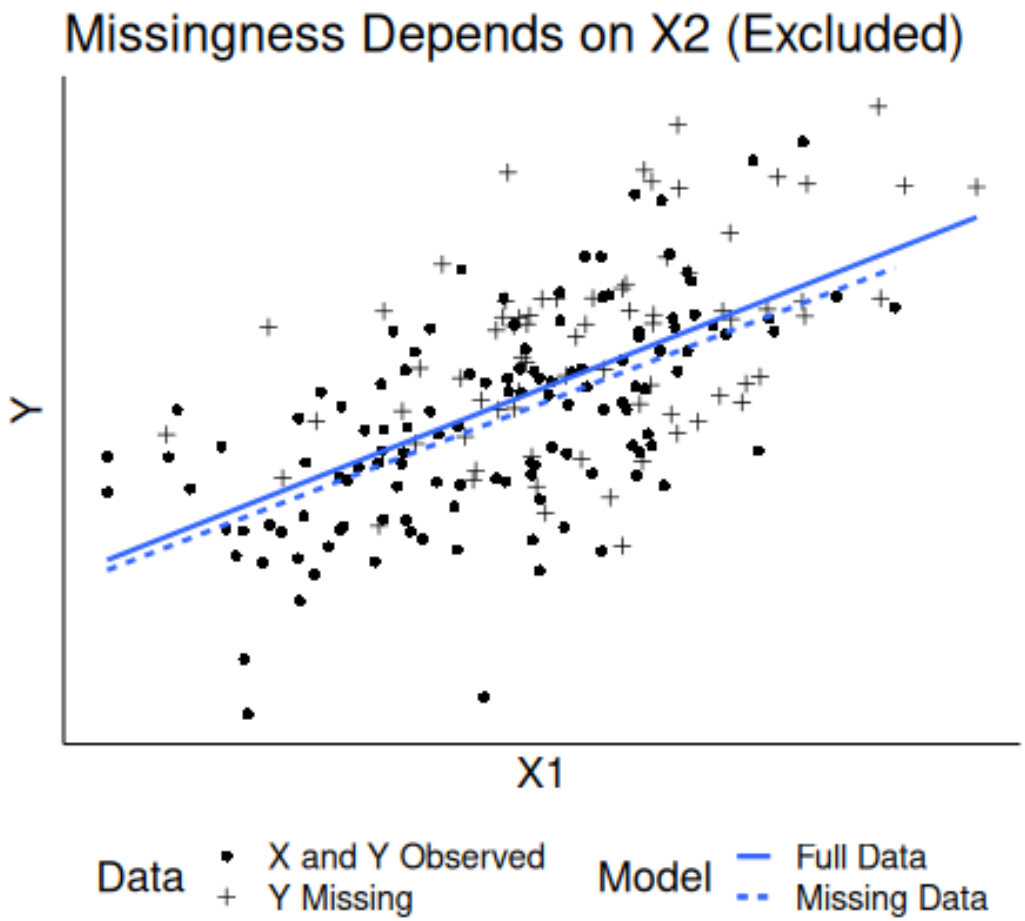
Listwise Deletion (Complete Case Analysis)

- MCAR: unbiased
- MAR where missingness depends on X-variable included in model: unbiased



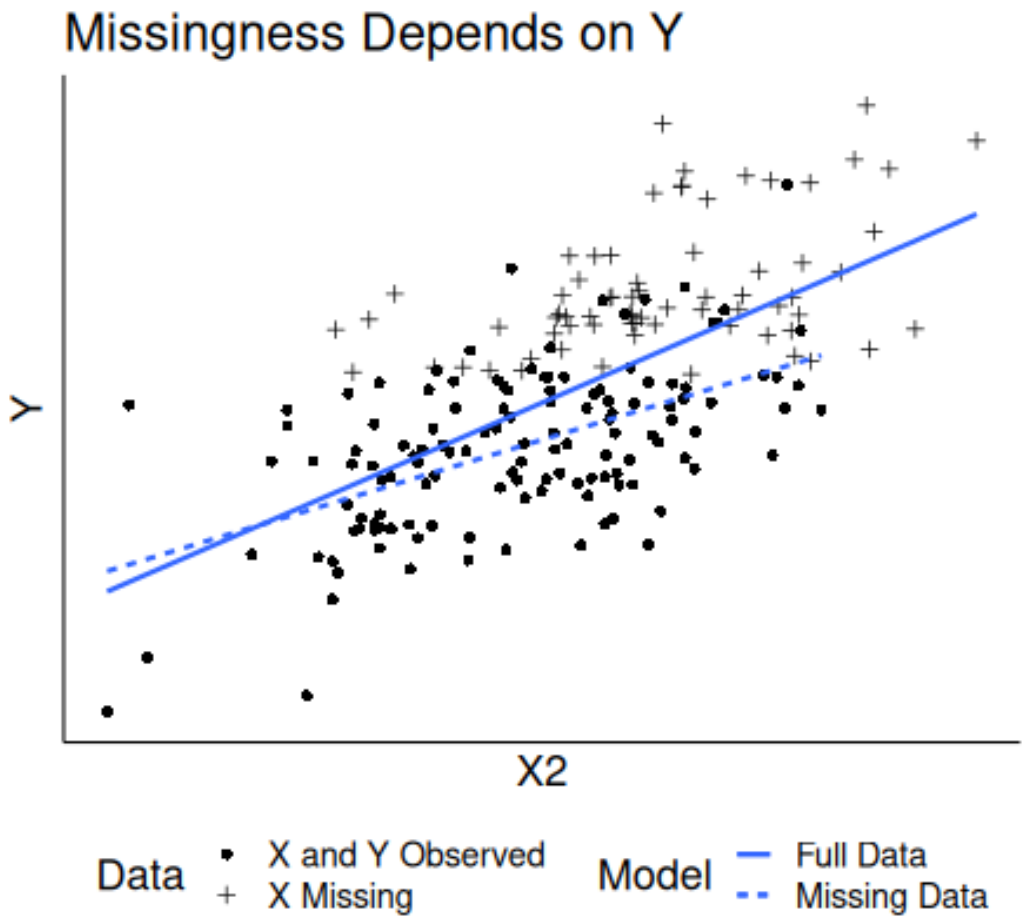
Listwise Deletion (Complete Case Analysis)

- MCAR: unbiased
- MAR where missingness depends on X-variable included in model: unbiased
- MAR where missingness depends on excluded X-variable: (potentially) biased



Listwise Deletion (Complete Case Analysis)

- MCAR: unbiased
- MAR where missingness depends on X-variable included in model: unbiased
- MAR where missingness depends on excluded X-variable: (potentially) biased
- MAR where missingness in X depends on (observed and included) Y-variable: biased



Listwise Deletion

Based on 1,000 simulated datasets of size n = 500 with 40% missingness:

Setting	Coef.	Bias	SD($\hat{\beta}$)	95% CI Coverage
Complete Data	β_1	0	0.044	95%
MCAR	β_1	0	0.058	94%
Missing $Y\|X_2$ (in model)	β_1	0	0.056	96%
Missing $Y\|X_2$ (not in model)	β_1	-0.06	0.057	83%
Missing $X_2\|Y$	β_1	-0.08	0.056	67%
Complete Data	β_2	0	0.041	97%
MCAR	β_2	0	0.054	97%
Missing $Y\|X_2$	β_2	0	0.064	95%
Missing $X_2\|Y$	β_2	-0.08	0.054	66%

Listwise Deletion (Complete Case Analysis)

- Effectively assumes those with complete data can be treated as a representative subsample
  - Valid if missingness is MCAR
  - Valid if missingness in Y depends only on (modeled) X
- Invalid if missingness (in X or Y) depends on value of Y

Listwise Deletion (Complete Case Analysis)

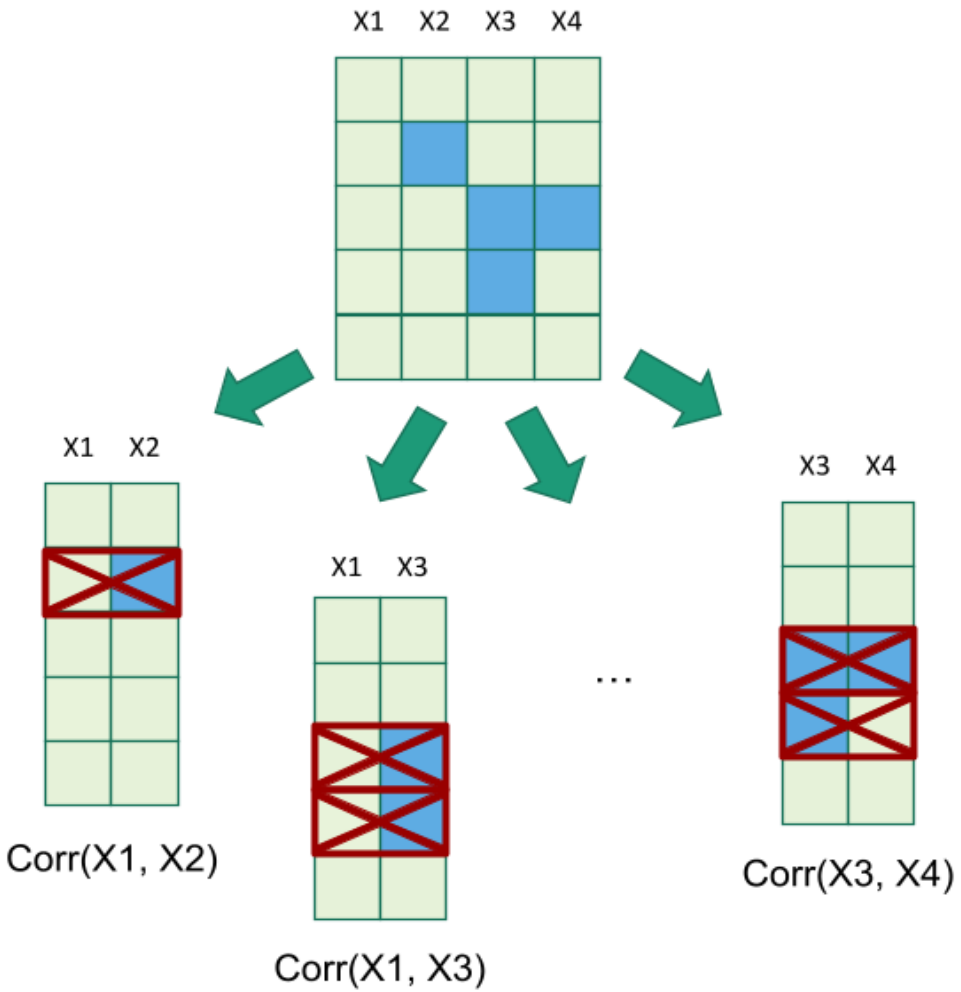
- Default in most statistical software (e.g., `lm()`)
- **Advantages**
  - Can be used for any downstream statistical analysis
  - No special computational methods required
  - **If MCAR**, then observed data are a random subsample of full data $\implies$ all estimates are just as valid in subsample
- **Disadvantages**
  - **Inefficient**: SEs are larger than both full data analysis and more sophisticated missing data methods
  - **If MAR**: estimates may be biased if probability of missing data in X depends on Y

Conventional Approaches

- Listwise deletion (complete case analysis)
- Pairwise deletion (available data analysis)
- Dummy variable adjustment (missing-indicator method)
- Last observation carried forward
- Single imputation methods

Pairwise Deletion (Available Data Analysis)

- **Cross-sectional study:** Remove subjects with missing data *on variables used for particular statistics*
  - E.g., for $Corr(X_1, X_2)$, use all cases with both X_1 and X_2



Pairwise Deletion (Available Data Analysis)

- **Cross-sectional study:** Remove subjects with missing data *on variables used for particular statistics*
  - E.g., for $Corr(X_1, X_2)$, use all cases with both X_1 and X_2 (`use="pairwise.complete.obs"` in R)

```
> library(psych)
> corr.test(nhanes.sub, use="pairwise.complete.obs")
Call:corr.test(x = nhanes.sub, use = "pairwise.complete.obs")
Correlation matrix
          TotChol  BMI Pulse BPSysAve
TotChol      1.00 0.11  0.04     0.15
BMI          0.11 1.00  0.15     0.32
Pulse        0.04 0.15  1.00     0.10
BPSysAve     0.15 0.32  0.10     1.00
Sample Size
          TotChol  BMI Pulse BPSysAve
TotChol     1503 1495  1430     1428
BMI         1495 1495  1425     1423
Pulse       1430 1425  1430     1428
BPSysAve    1428 1423  1428     1428
```

Pairwise Deletion (Available Data Analysis)

- **Longitudinal study:** Remove *only the time points with missing data* (on variables used in analysis)

ID	Time	X_1	X_2	Y
1	1	4.25	1	8.53
1	2	3.61	1	
1	3	2.98	1	
2	1	4.93	0	5.58
2	2	4.62	0	7.21
2	3	5.84	0	11.64
3	1	4.28	0	6.42
3	2	5.63	0	10.22
3	3	4.64	0	

$\rightarrow$

ID	Time	X_1	X_2	Y
1	1	4.25	1	8.53
2	1	4.93	0	5.58
2	2	4.62	0	7.21
2	3	5.84	0	11.64
3	1	4.28	0	6.42
3	2	5.63	0	10.22

Pairwise Deletion - Inference for β_1

Based on 1,000 simulated datasets of size n $= 500$ with 40% missingness:

Setting	$\rho(X_1, X_2)$	Method	Bias	SD($\hat{\beta}$)	95% CI Coverage
Complete Data	0.8	Listwise	0	0.063	95%
	0.8	Pairwise	0	0.063	95%
MCAR	0.8	Listwise	0.01	0.082	95%
	0.8	Pairwise	0	0.105	94%
Missing $Y \mid X_2$	0.8	Listwise	0	0.080	96%
	0.8	Pairwise	0	0.103	95%
Missing $X_2 \mid Y$	0.8	Listwise	-0.07	0.077	85%
	0.8	Pairwise	0.32	0.080	5%

Pairwise Deletion - Inference for β_2

Based on 1,000 simulated datasets of size n = 500 with 40% missingness:

Setting	$\rho(X_1, X_2)$	Method	Bias	SD($\hat{\beta}$)	95% CI Coverage
Complete Data	0.8	Listwise	0	0.060	97%
	0.8	Pairwise	0	0.060	97%
MCAR	0.8	Listwise	0	0.079	96%
	0.8	Pairwise	0	0.099	95%
Missing $Y \mid X_2$	0.8	Listwise	0	0.084	97%
	0.8	Pairwise	-0.25	0.102	34%
Missing $X_2 \mid Y$	0.8	Listwise	-0.08	0.076	84%
	0.8	Pairwise	-0.38	0.098	4%

Pairwise Deletion (Available Data Analysis)

- **Advantages**
  - Can be used for any downstream statistical analysis
  - **If MCAR**, then estimates are still consistent
  - More efficient than complete case analysis **if** correlation among x-variables is low (and typically for longitudinal settings)

- **Disadvantages**
  - Estimated standard errors are typically **not** unbiased, even with MCAR
    - It's hard to determine the effective sample size/df
    - Some `R` packages (e.g., `regtools`) use bootstrap standard errors instead - valid SE estimation
  - **If MAR**, then estimates may be biased

- Takeaway: not clearly better than complete case analysis (at least as used in cross-sectional studies)

Conventional Approaches

- Listwise deletion (complete case analysis)
- Pairwise deletion (available data analysis)
- Dummy variable adjustment (missing-indicator method)
- Last observation carried forward
- Single imputation methods

Dummy Variable Adjustment

Suppose X_2 has missing data:

- Define missingness indicator $R_2 = I(X_2 \text{ missing})$
- Use R_2 and $X_2(1 - R_2)$ as predictors in model
- E.g., if our desired model is $E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$:

ID	X_1	X_2	Y
1	4.67	0	5.16
2	3.35		4.73
3	1.63		3.25
4	3.96	1	4.96
5	0.13	0	-0.6
6	1.21	1	4.26

$\rightarrow$

ID	X_1	$X_2(1 - R_2)$	R_2	Y
1	4.67	0	0	5.16
2	3.35	0	1	4.73
3	1.63	0	1	3.25
4	3.96	1	0	4.96
5	0.13	0	0	-0.6
6	1.21	1	0	4.26

- Then fit model $E(Y|X, R) = \beta_0 + \beta_1 X_1 + \beta_2 X_2(1 - R_2) + \beta_3 R_2$

Dummy Variable Adjustment

Based on 1,000 simulated datasets of size n = 500, $\rho(X_1, X_2) = 0.5$:

Setting	Coef.	Bias	SD($\hat{\beta}$)	95% CI Coverage
Complete Data	β_1	0	0.044	95%
MCAR	β_1	0.13	0.044	15%
Missing $X_2 \mid X_2$	β_1	0.05	0.044	83%
Missing $X_2 \mid Y$	β_1	-0.11	0.039	22%
Complete Data	β_2	0	0.041	97%
MCAR	β_2	-0.07	0.054	84%
Missing $X_2 \mid X_2$	β_2	-0.02	0.062	94%
Missing $X_2 \mid Y$	β_2	-0.07	0.052	67%

Dummy Variable Adjustment - Modification

A modification of the previous model:

$$E(Y|X,R) = \beta_0 + \beta_1 X_1 + \beta_2 X_2(1-R_2) + \beta_3 R_2$$

$$\rightarrow$$

$$E(Y|X,R) = \beta_0 + \beta_1 X_1(1-R_2) + \beta_2 X_2(1-R_2) + \beta_3 R_2 + \beta_4 X_1 R_2$$

ID	$X_1(1-R_2)$	X_1R_2	$X_2(1-R_2)$	R_2	Y
1	4.67	0	0	0	5.16
2	0	3.35	0	1	4.73
3	0	1.63	0	1	3.25
4	3.96	0	1	0	4.96
5	0.13	0	0	0	-0.6
6	1.21	0	1	0	4.26

Dummy Variable Adjustment - Modification

Based on 1,000 simulated datasets of size n = 500, $\rho(X_1, X_2) = 0.5^{\star}$:

Setting	Coef.	Bias	SD($\hat{\beta}$)	95% CI Coverage
Complete Data	β_1	0	0.08	95%
MCAR	β_1	0	0.11	100%
Missing $X_2 \vert X_2$	β_1	0	0.11	100%
Missing $X_2 \vert Y$	β_1	-0.02	0.11	100%
Complete Data	β_2	0	0.042	95%
MCAR	β_2	0	0.053	100%
Missing $X_2 \vert X_2$	β_2	0	0.062	100%
Missing $X_2 \vert Y$	β_2	-0.04	0.061	100%

*the other parameters (e.g, β_1) are different than previous ones, to emphasize the coverage issues

Stratification Methods

Suppose again X_2 has missing data and is a categorical variable (e.g., treatment indicator, age category):

- Define new strata to include those missing X_2

ID	X_1	X_2	Y
1	4.67	0	5.16
2	3.35		4.73
3	1.63		3.25
4	3.96	1	4.96
5	0.13	0	-0.6
6	1.21	1	4.26

$\rightarrow$

ID	X_1	$I(X_2$ obs, $X_2 = 1)$	$I(X_2$ miss$)$	Y
1	4.67	0	0	5.16
2	3.35	0	1	4.73
3	1.63	0	1	3.25
4	3.96	1	0	4.96
5	0.13	0	0	-0.6
6	1.21	1	0	4.26

- Then fit model
  $E(Y|X,R) = \beta_0 + \beta_1 X_1 + \beta_2 I(X_2 \text{ obs}, X_2 = 1) + \beta_3 I(X_2 \text{ miss})$
- Modification: allow for different β_1 for missing-data stratum:

$$E(Y|X,R) = \beta_0 + \beta_1 X_1 I(X_2 \text{ obs}) + \beta_2 I(X_2 \text{ obs}, X_2 = 1) + \beta_3 I(X_2 \text{ miss}) + \beta_4 X_1 I(X_2 \text{ miss})$$

Dummy Variable Adjustment and Stratification Methods

- **Advantages**
  - Uses all available information (except cases with missing outcomes)
- **Disadvantages**
  - Can produce biased coefficient estimates (even under MCAR)
  - Modifications can result in unbiased coefficient estimates, but overestimate residual variance leading to standard errors that are too large
- Takeaway: Not recommended.

Conventional Approaches

- Listwise deletion (complete case analysis)
- Pairwise deletion (available data analysis)
- Dummy variable adjustment (missing-indicator method)
- Last observation carried forward
- Single imputation methods

Last Observation Carried Forward

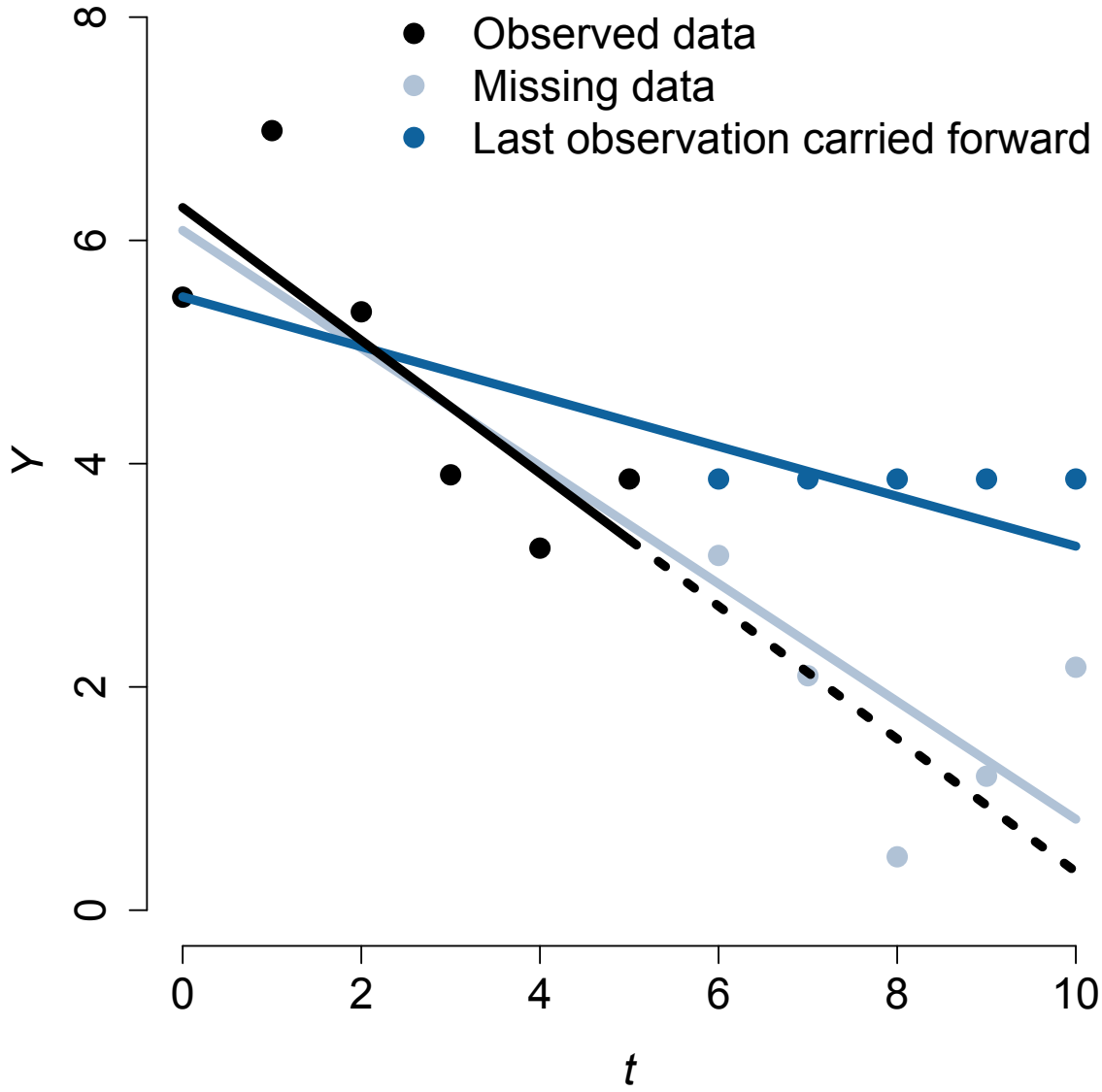
Only for longitudinal studies: replace missing values with subject's last observed value

ID	Time	X_1	X_2	Y
1	1	4.25	1	8.53
1	2	3.61	1	
1	3	2.98	1	
2	1	4.93	0	5.58
2	2	4.62	0	7.21
2	3	5.84	0	11.64
3	1	4.28	0	6.42
3	2	5.63	0	10.22
3	3	4.64	0	

$\rightarrow$

ID	Time	X_1	X_2	Y
1	1	4.25	1	8.53
1	2	3.61	1	8.53
1	3	2.98	1	8.53
2	1	4.93	0	5.58
2	2	4.62	0	7.21
2	3	5.84	0	11.64
3	1	4.28	0	6.42
3	2	5.63	0	10.22
3	3	4.64	0	10.22

Last Observation Carried Forward



Last Observation Carried Forward

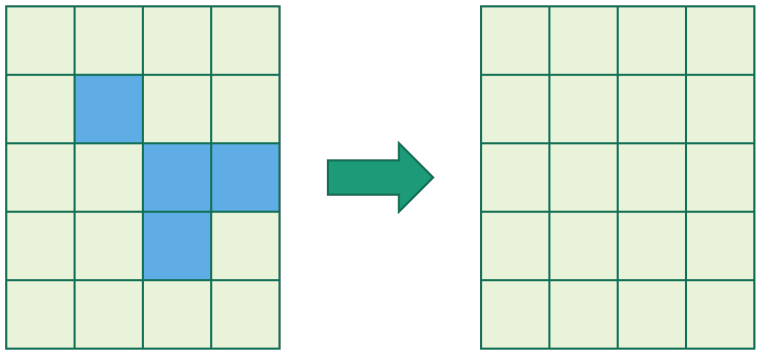
- **Advantages**
  - Might be justifiable under a "last observation analysis" framework – answers a different question
- **Disadvantages**
  - May result in serious bias in either direction (even under MCAR)
  - May result in anti-conservative p-values; variance is understated
  - Has been thoroughly repudiated, but still a standard method used by the pharmaceutical industry and appears in published articles
- Takeaway: Not recommended.

Conventional Approaches

- Listwise deletion (complete case analysis)
- Pairwise deletion (available data analysis)
- Dummy variable adjustment (missing-indicator method)
- Last observation carried forward
- Single imputation methods

Single Imputation

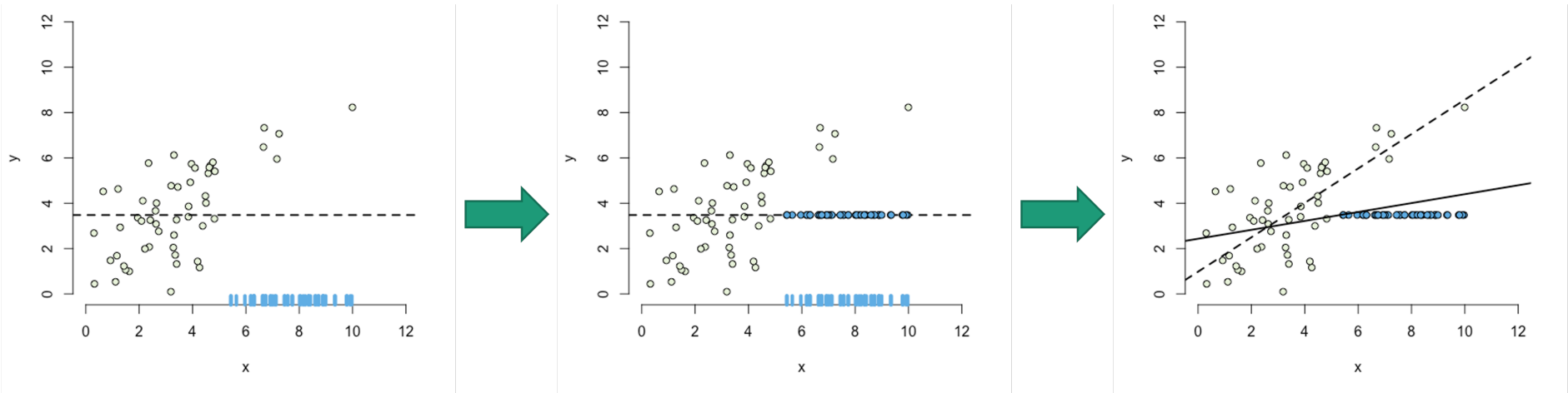
- Imputation = "filling in" missing data
- *Single* imputation: create just one completed dataset
  - Later, we'll talk about *multiple* imputation



- There are a variety of ways to do this
  - Mean imputation
  - Regression imputation
  - Stochastic regression imputation
  - Hot deck imputation

Single Imputation: (Unconditional) Mean Imputation

Mean imputation: Replace missing values with the mean of that variable



Single Imputation: (Unconditional) Mean Imputation

Based on 1,000 simulated datasets of size n = 500, $\rho(X_1, X_2) = 0.5$:

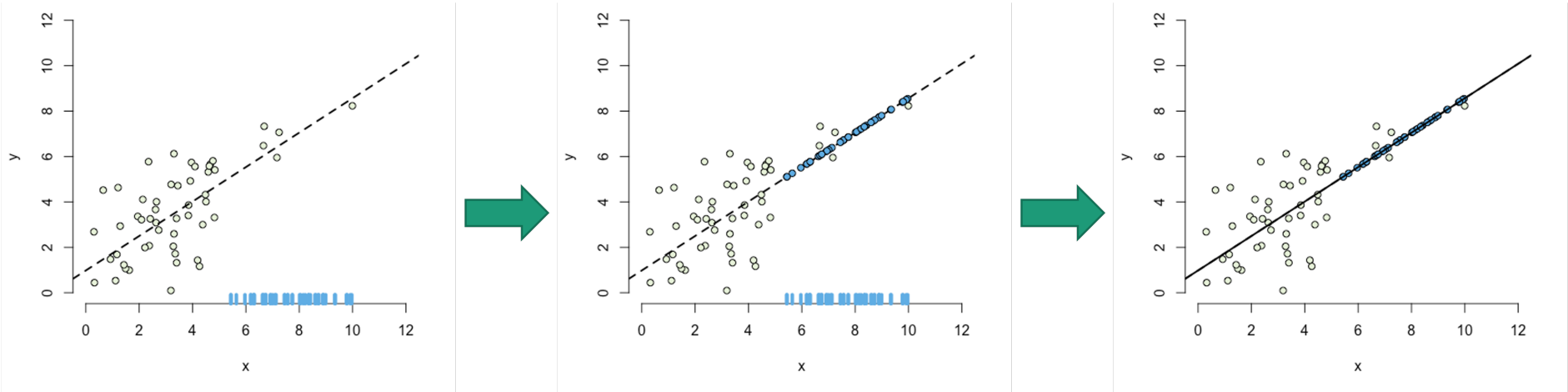
Setting	Coef.	Bias	SD($\hat{\beta}$)	95% CI Coverage
Complete Data	β_1	0	0.044	95%
MCAR	β_1	-0.20	0.045	1%
Missing $Y \mid X_2$	β_1	-0.20	0.044	1%
Missing $X_2 \mid Y$	β_1	0.18	0.044	2%
Complete Data	β_2	0	0.041	97%
MCAR	β_2	-0.20	0.042	0%
Missing $Y \mid X_2$	β_2	-0.328	0.041	0%
Missing $X_2 \mid Y$	β_2	-0.020	0.054	7%

Single Imputation: (Unconditional) Mean Imputation

- **Advantages**
  - Easy/quick to implement
- **Disadvantages**
  - Biases estimate of mean when data are not MCAR
  - Underestimates variance
  - Distorts relationships between variables (i.e., regression coefficients may be biased even under MCAR)
- Takeaway: Not recommended.

Single Imputation: Regression Imputation

Regression imputation: Replace missing values with the predictions from a regression model



Single Imputation: Regression Imputation

Based on 1,000 simulated datasets of size n = 500, $\rho(X_1, X_2) = 0.5$:

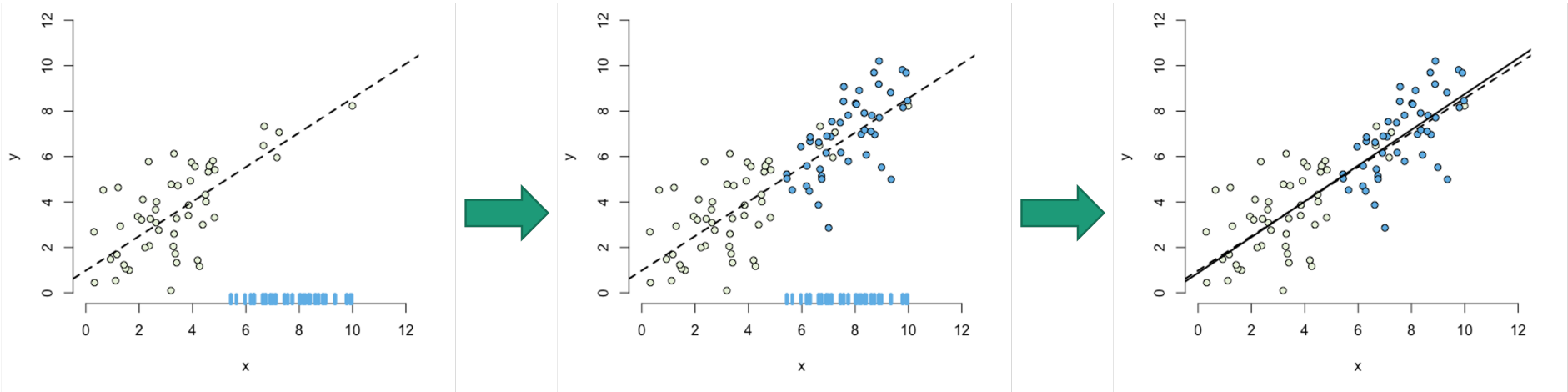
Setting	Coef.	Bias	SD($\hat{\beta}$)	95% CI Coverage
Complete Data	β_1	0	0.044	95%
MCAR	β_1	0	0.058	77%
Missing $Y \mid X_2$	β_1	0	0.056	79%
Missing $X_2 \mid Y$	β_1	-0.12	0.061	30%
Complete Data	β_2	0	0.041	97%
MCAR	β_2	0	0.054	80%
Missing $Y \mid X_2$	β_2	0	0.064	72%
Missing $X_2 \mid Y$	β_2	0.23	0.067	5%

Single Imputation: Regression Imputation

- **Advantages**
  - Easy/quick to implement
- **Disadvantages**
  - Underestimates variance
  - Correlations biased upwards
  - Deceptive! Imputations too good to be true
- Takeaway: Not recommended.

Single Imputation: Stochastic Regression Imputation

Stochastic regression imputation: Replace missing values with the predictions from a regression model plus random noise



Single Imputation: Stochastic Regression Imputation

Based on 1,000 simulated datasets of size n = 500, $\rho(X_1, X_2) = 0.5$:

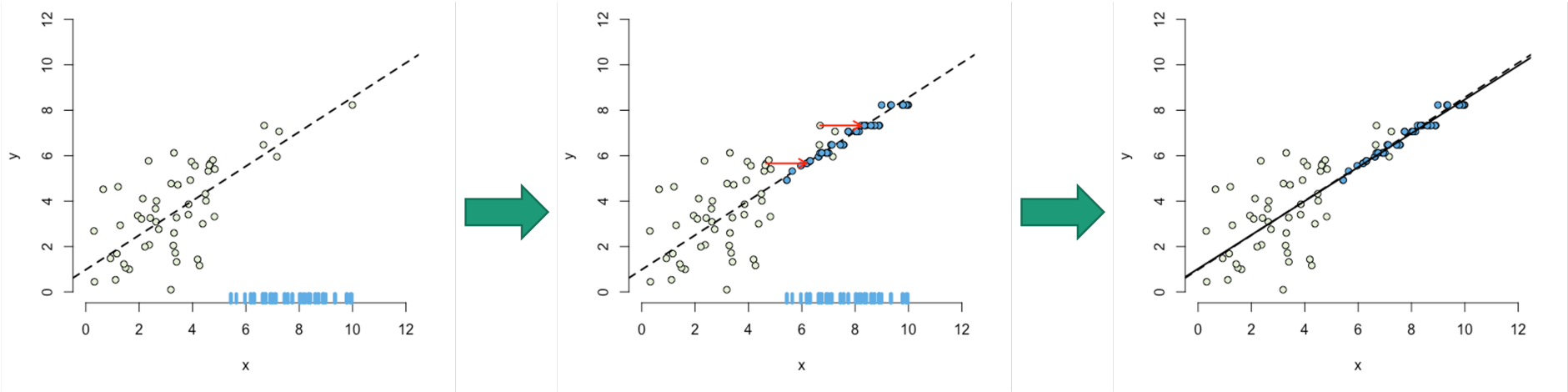
Setting	Coef.	Bias	SD($\hat{\beta}$)	95% CI Coverage
Complete Data	β_1	0	0.044	95%
MCAR	β_1	0	0.065	83%
Missing $Y \mid X_2$	β_1	0	0.062	84%
Missing $X_2 \mid Y$	β_1	0	0.055	89%
Complete Data	β_2	0	0.041	97%
MCAR	β_2	0	0.061	85%
Missing $Y \mid X_2$	β_2	0	0.069	80%
Missing $X_2 \mid Y$	β_2	0	0.060	86%

Single Imputation: Stochastic Regression Imputation

- **Advantages**
  - Quick to implement
  - Unbiased for mean, regression coefficients, correlation under MAR
- **Disadvantages**
  - Standard error still underestimated
- Takeaway: Not recommended for inference
  - If sharing one imputed dataset, this is (arguably) the best approach

Single Imputation: Hot Deck

Hot deck imputation: Replace missing values with observed values from "similar" respondents (example: predictive mean matching)



Single Imputation: Hot Deck Imputation

Based on 1,000 simulated datasets of size n = 500, $\rho(X_1, X_2) = 0.5$:

Setting	Coef.	Bias	SD($\hat{\beta}$)	95% CI Coverage
Complete Data	β_1	0	0.044	95%
MCAR	β_1	0	0.068	82%
Missing $Y \mid X_2$	β_1	-0.01	0.070	79%
Missing $X_2 \mid Y$	β_1	0.01	0.062	84%
Complete Data	β_2	0	0.041	97%
MCAR	β_2	-0.01	0.064	84%
Missing $Y \mid X_2$	β_2	-0.02	0.080	73%
Missing $X_2 \mid Y$	β_2	-0.01	0.071	78%

Single Imputation: Hot Deck

- **Advantages**
  - Imputations are based on observed values, so are realistic
  - Less sensitive to imputation model misspecification
- **Disadvantages**
  - Results can be sensitive to number of candidate donors
  - May not perform well in small datasets
  - Standard error still underestimated

Conventional Approaches: Takeaways

All the common methods for salvaging information from cases with missing data typically make things worse: They introduce substantial bias, make the analysis more sensitive to departures from MCAR, or yield standard error estimates that are incorrect (usually too low). –Paul Allison

The takeaway message [. . .] is that ad hoc approaches to accounting for missing data that are not based on a formal, principled statistical framework are likely to lead to erroneous inference. –Marie Davidian

Conventional Approaches: Takeaways

- Software default of listwise deletion is the least dangerous of the conventional methods
  - Valid if data are MCAR
  - Valid if data are MAR and missingness depends only on modeled X
- Most of the conventional and naive methods can lead to biased estimates (in either direction) and/or invalid inference
- We can do better!

Overview

Maximum Likelihood Overview

Begin with an assumed model for the data. For example, $Y \sim N(\mu, \sigma^2)$.

Maximum likelihood (ML) is a general approach for estimating the parameters in our assumed model (e.g. estimating μ and σ^2)

⋆ General idea: choose values that make the data observed most likely

Maximum Likelihood Overview

Suppose we have **data y** and are interested in estimating a **parameter** θ

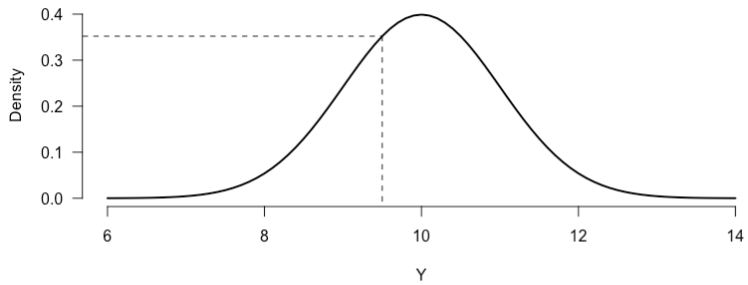
- Data assumed to be generated by a model described by a **probability density function** $f(\mathbf{y}|\boldsymbol{\theta})$
  - Describes the relative probability of obtaining a sample **y** for given θ
- The **likelihood function** $L(\theta|\mathbf{y}) = f(\mathbf{y}|\theta)$ is function of the parameter θ for given data **y**
- A **maximum likelihood (ML) estimate** of θ is a value that maximizes the likelihood $L(\theta|\mathbf{y})$ (or equivalently log-likelihood $l(\theta|\mathbf{y}) = \log L(\theta|\mathbf{y})$)

An Example

- Assume $Y \sim N(\mu, 1)$
  - Eventually, our goal will be to estimate the mean μ
- The **probability density function** for a single observation is:

$$f(y|\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu)^2}$$

- Let's start by supposing that $\mu = 10$



- For a sample of size 100, we observe $(Y_1, Y_2, \ldots, Y_{100})$. The **joint probability density** is:

$$f(\mathbf{y}|\mu = 10) = \prod_{i=1}^{100} f(y_i|\mu = 10)$$

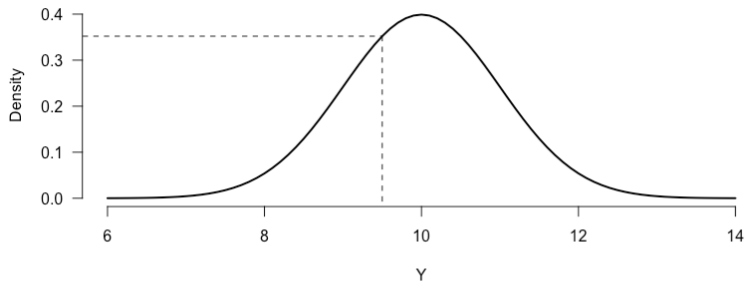
y_i	$f(y_i\|\mu = 10)$
9.5	0.35
10.5	0.35
8.5	0.13
⋮	⋮
11.5	0.13
$f(\mathbf{y}\|\mu = 10)$	5.6×10^{-58}

An Example

- Assume $Y \sim N(\mu, 1)$
  - Eventually, our goal will be to estimate the mean μ
- The **probability density function** for a single observation is:

$$f(y|\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu)^2}$$

- Let's start by supposing that $\mu = 10$



- For a sample of size 100, we observe $(Y_1, Y_2, \ldots, Y_{100})$. The **joint probability density** is:

$$f(\mathbf{y}|\mu = 10) = \prod_{i=1}^{100} f(y_i|\mu = 10)$$

y_i	$f(y_i\|\mu = 10)$
9.5	0.35
10.5	0.35
8.5	0.13
$\vdots$	$\vdots$
11.5	0.13
$f(\mathbf{y}\|\mu = 10)$	5.6×10^{-58}

An Example

Goal: estimate μ

- Idea: find value of μ that makes our observed data most probable
- The **likelihood function** is:

$$L(\mu|\mathbf{y}) = f(\mathbf{y}|\mu)$$

- It's easier to work with logs $\rightarrow$ **log-likelihood**:

$$l(\mu|\mathbf{y}) = \log L(\mu|\mathbf{y})$$

- How to find value of μ that maximizes the likelihood (equivalently log-likelihood)?

y_i	$\mu = 9$	$\mu = 10$	$\mu = 11$
9.5	-1.04	-1.04	-2.04
10.5	-2.04	-1.04	-1.04
8.5	-1.04	-2.04	-4.04
⋮	⋮	⋮	⋮
11.5	-4.04	-2.04	-1.04
$l(\mu\|\mathbf{y})$	-182	-132	-182

An Example

Goal: estimate μ

- Idea: find value of μ that makes our observed data most probable
- The **likelihood function** is:

$$L(\mu|\mathbf{y}) = f(\mathbf{y}|\mu)$$

- It's easier to work with logs $\rightarrow$ **log-likelihood**:

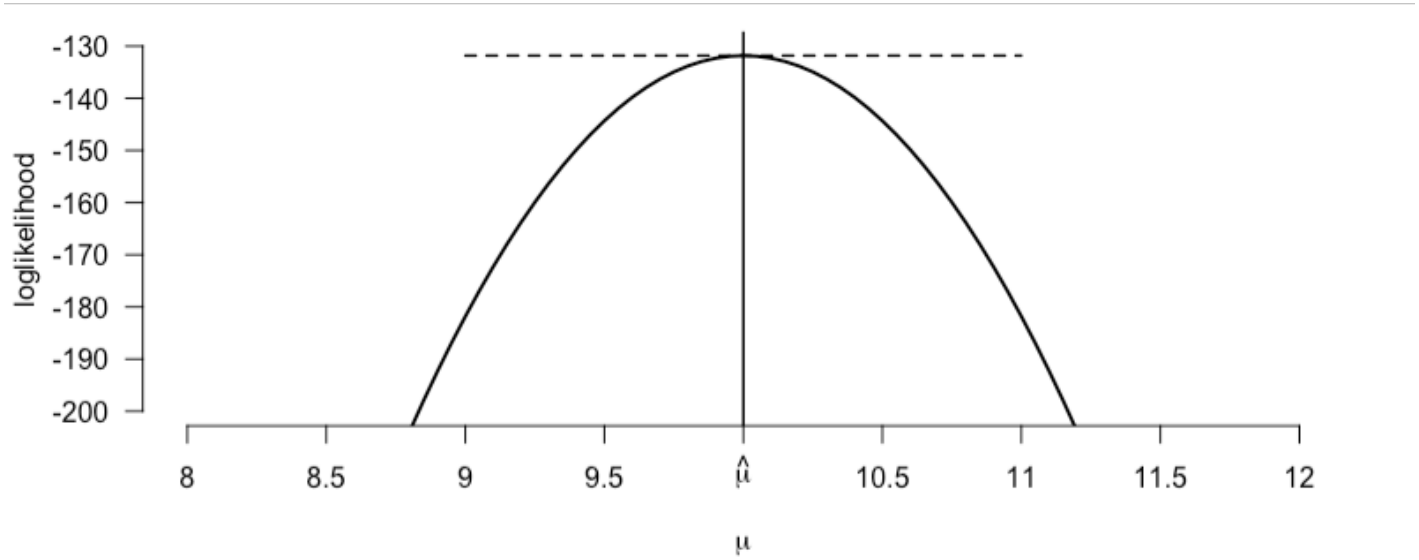
$$l(\mu|\mathbf{y}) = \log L(\mu|\mathbf{y})$$

- How to find value of μ that maximizes the likelihood (equivalently log-likelihood)?

y_i	$\mu = 9$	$\mu = 10$	$\mu = 11$
9.5	-1.04	-1.04	-2.04
10.5	-2.04	-1.04	-1.04
8.5	-1.04	-2.04	-4.04
⋮	⋮	⋮	⋮
11.5	-4.04	-2.04	-1.04
$l(\mu\|\mathbf{y})$	-182	-132	-182

An Example

- Here, we can find the value of μ that maximizes the log-likelihood by finding the derivative and seeing where that equals 0



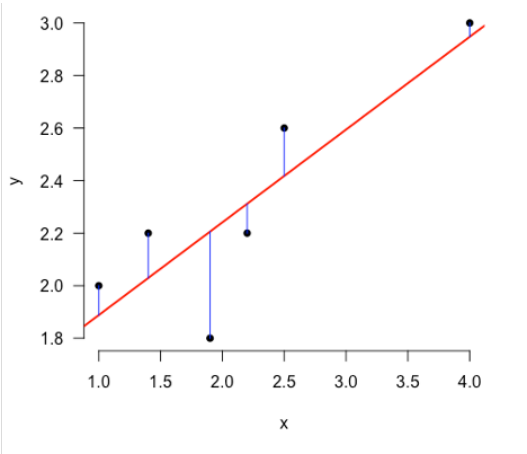
- In this particular example, there is a closed form solution for $\hat{\mu}$ (the sample mean!)

Another Example: Linear Regression

Ordinary least squares approach equivalent to ML when error term normally distributed:

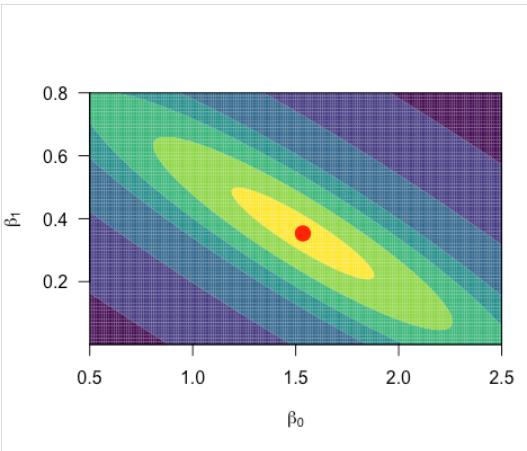
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2)$$

$$L(\theta|\mathbf{y}) = \prod_{i=1}^{n} f(y_i|\theta) = \prod_{i=1}^{n} N(y_i; \mu_i = \beta_0 + \beta_1 x_i, \sigma^2)$$



Maximizing likelihood L is equivalent to *minimizing* negative log-likelihood l:

$$l(\mathbf{y}|\theta) = \sum_{i=1}^{n} \log\left(N(y_i; \mu_i = \beta_0 + \beta_1 x_i, \sigma^2)\right)$$



$$\hat{\beta}_0, \hat{\beta}_1 = \text{values that } \textit{minimize} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Another Example: Linear Regression

To find the *minimum*, we take the derivative, set equal to 0 and solve. This leads us to the following equations for our MLEs:

$$\hat{\beta}_1 = \text{corr}(x, y) \times \frac{\text{SD}(y)}{\text{SD}(x)}$$

$$\hat{\beta}_0 = \text{mean}(y) - \hat{\beta}_1 \times \text{mean}(x)$$

While this works nicely in this setting, this is not always the case. When no closed form solution can be found, *iterative* techniques such as the Newton-Raphson algorithm or Fisher Scoring method can be used to obtain the MLEs.

Properties of ML estimators

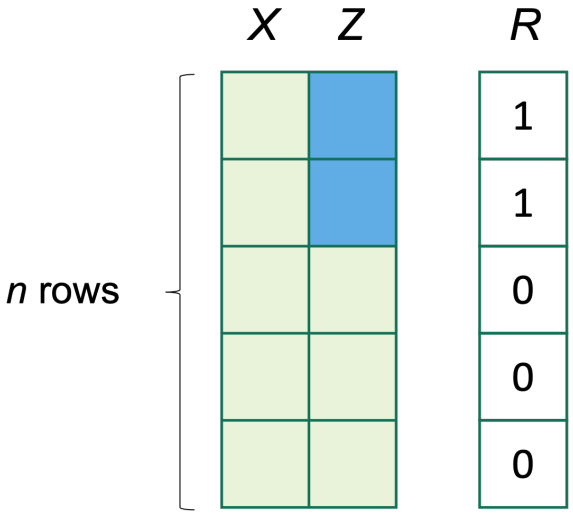
In general (under certain conditions), ML estimators are:

- **Consistent**: approximately unbiased in large samples (hone in on the right thing)
- **Asymptotically efficient**: smallest variance among all consistent estimators in large samples
- **Asymptotically normal**: estimator has approximately a normal distribution
  - Approximation improves as sample size increases
  - Can use normal distribution to construct confidence intervals (e.g. 1.96) or obtain p-values

Now With Missingness

- Same idea: find value(s) of parameter(s) with largest log-likelihood
- Some adjustments:
  1. Likelihood contributions from individuals with missing values
  2. Standard error computations tend to be trickier
  3. Usually requires iterative optimization algorithms

From Complete To Missing



Now suppose that we are missing several Z values. That is, our observed data are $\mathbf{x}$, $\mathbf{z}^{(obs)}$, and $\mathbf{r}$.

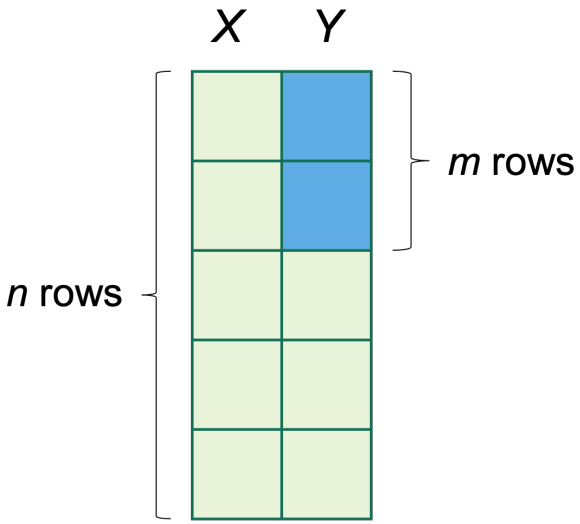
The full likelihood is

$$L(\theta, \phi|\mathbf{x}, \mathbf{z}^{(obs)}, \mathbf{r}) = \int f(\mathbf{x}, \mathbf{z}^{(obs)}, \mathbf{z}^{(mis)}|\theta) \underbrace{f(\mathbf{r}|\mathbf{z}^{(obs)}, \mathbf{z}^{(mis)}, \mathbf{x}, \phi)}_{\text{If MAR} = f(\mathbf{r}|\mathbf{z}^{(obs)}, \mathbf{x}, \phi)} d\mathbf{z}^{(mis)}$$

Assuming the missing data mechanism is **ignorable** (slide 21), the likelihoood simplifies to

$$L(\theta|\mathbf{x}, \mathbf{z}^{(obs)}) = \int f(\mathbf{x}, \mathbf{z}^{(obs)}, \mathbf{z}^{(mis)}|\theta) d\mathbf{z}^{(mis)}$$

Linear Regression with Missing Response



Goal: estimate β_0 and β_1

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2)$$

where Y is missing for m observations. Define $\theta = (\beta_0, \beta_1, \sigma^2)$. Assume missingness depends on X (i.e., MAR).

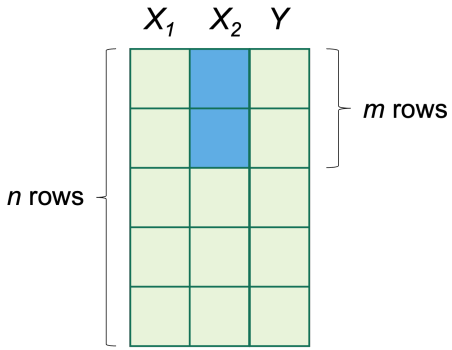
The observed-data likelihood ignoring the missing data mechanism is:

$$L(\theta|\mathbf{x}, \mathbf{y}^{(obs)}) = \prod_{i=1}^{m} f(x_i|\theta) \prod_{i=m+1}^{n} f(x_i, y_i|\theta)$$

$$\propto \prod_{i=m+1}^{n} \underbrace{f(y_i|x_i, \theta)}_{N(\beta_0+\beta_1 x_i, \sigma^2)}$$

Conclusion: reduces to listwise deletion here

Multiple Linear Regression with Missing Covariate



Suppose we are interested in estimating β_0, β_1, and β_2:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2)$$

where X_{2i} is missing for m observations. We'll also assume $X_2|X_1 \sim N(\alpha_0 + \alpha_1 X_1, \tau^2)$. Define $\theta = (\beta_0, \beta_1, \beta_2, \sigma^2, \alpha_0, \alpha_1, \tau^2)$.

The observed-data likelihood assuming ignorability is:

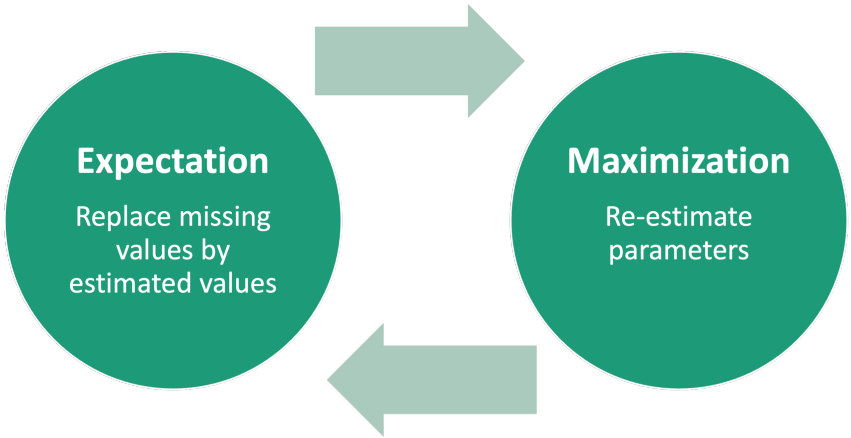
$$L(\theta|\mathbf{x}_1, \mathbf{x}_2^{(obs)}, \mathbf{y}) = \prod_{i=1}^{m} f(x_{1i}, y_i|\theta) \prod_{i=m+1}^{n} f(x_{1i}, x_{2i}, y_i|\theta)$$

$$\propto \prod_{i=1}^{m} f(y_i|x_{1i}, \theta) \prod_{i=m+1}^{n} \underbrace{f(y_i|x_{1i}, x_{2i}, \theta)}_{N(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}, \sigma^2)} \underbrace{f(x_{2i}|x_{1i}, \theta)}_{N(\alpha_0+\alpha_1 x_{1i}, \tau^2)}$$

The EM Algorithm

The **Expectation-Maximization (EM) algorithm** is a popular technique for general problems involving missing data[1]. Motivation for the EM algorithm:

- Complete-data likelihood (slide 91) is often easier to work with than the observed-data likelihood
- Do not need to compute second derivatives of the log-likelihood (required for other iterative methods)



[1] *Aside*: in this special case, we could directly derive $f(y_i|x_{1i}, \theta)$ and then maximize the observed-data likelihood

The EM Algorithm

0. Begin with initial estimate of parameters, $\theta^{(0)}$

For iteration t:

1. **E-step**: compute expectation of complete data log-likelihood given observed data and current parameter estimates

$$E\left[l(\theta|\mathbf{x}, \mathbf{z}^{(obs)}, \mathbf{z}^{(mis)})|\mathbf{x}, \mathbf{z}^{(obs)}, \theta = \theta^{(t-1)}\right]$$

2. **M-step**: maximize this expected log-likelihood to find $\theta^{(t)}$

Repeat until estimates converge

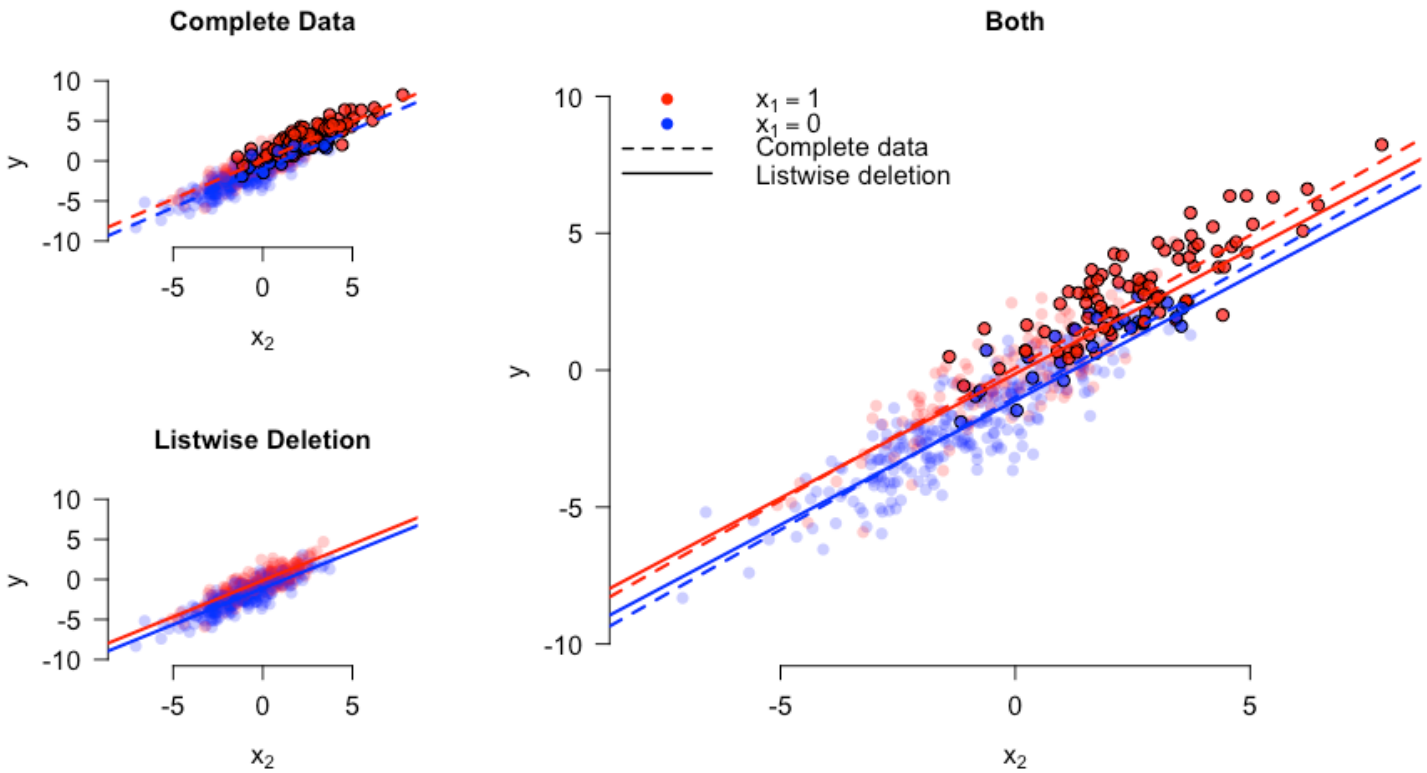
Dempster et al. (1977)

Multiple Linear Regression with Missing Covariate

Interested in estimating β_0, β_1, and β_2:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2)$$

where X_{2i} is missing for m observations, and missingness depends on Y.



Applying the EM Algorithm*

To implement the **E-step** of the algorithm, it turns out that we need $E[X_{2i}|\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2^{(obs)}, \theta = \theta^{(t-1)}]$ and $E[X_{2i}^2|\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2^{(obs)}, \theta = \theta^{(t-1)}]$

Observed data:

ID	y	x_1	x_2
1	0.29	0	NA
2	-0.95	0	NA
⋮	⋮	⋮	⋮
110	3.29	1	NA
111	-2.49	0	-1.52
⋮	⋮	⋮	⋮
500	-2.93	0	-0.51

$$\theta^{(0)} = \begin{pmatrix} \underbrace{0}_{\beta_0^{(0)}}, \underbrace{0}_{\beta_1^{(0)}}, \underbrace{0}_{\beta_2^{(0)}}, \underbrace{1}_{\sigma^{2(0)}}, \underbrace{0}_{\alpha_0^{(0)}}, \underbrace{0}_{\alpha_1^{(0)}}, \underbrace{1}_{\tau^{2(0)}} \end{pmatrix}$$

First iteration:

ID	y	x_1	$E[X_2]$	$E[X_2^2]$
1	0.29	0	0	1
2	-0.95	0	0	1
⋮	⋮	⋮	⋮	⋮
110	3.29	1	0	1
111	-2.49	0	-1.52	2.30
⋮	⋮	⋮	⋮	⋮
500	-2.93	0	-0.51	0.261

Applying the EM Algorithm*

In the **M-step**, we obtain new estimates of the parameters by maximizing the expected log likelihood and obtain $\theta^{(1)}$. We continue to iterate between these steps until convergence:

Second iteration:

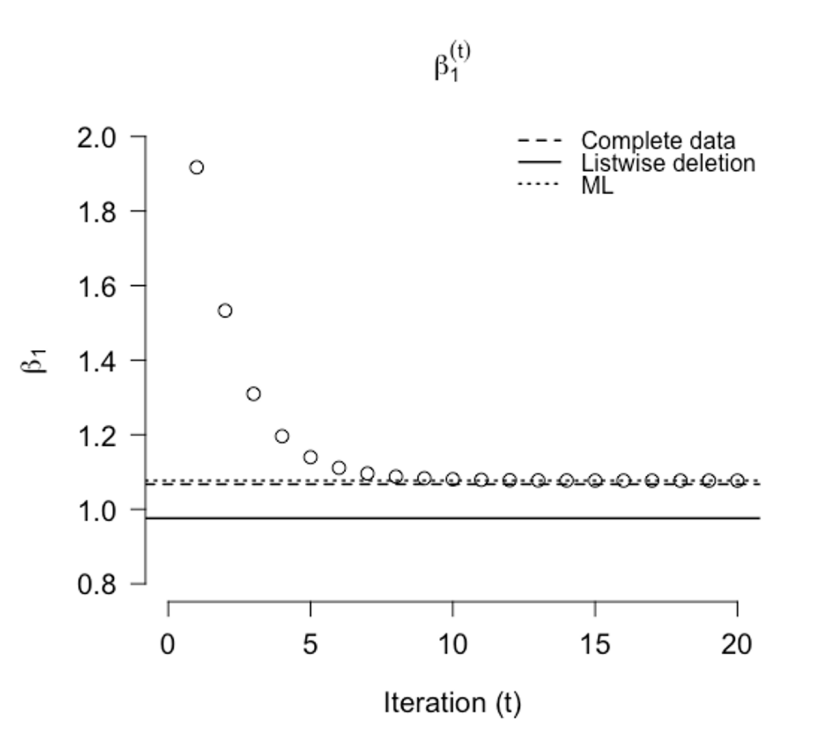
ID	y	x_1	$E[X_2]$	$E[X_2^2]$
1	0.29	0	0.11	1.58
2	-0.95	0	-0.53	1.86
⋮	⋮	⋮	⋮	⋮
110	3.29	1	1.14	2.88
111	-2.49	0	-1.52	2.30
⋮	⋮	⋮	⋮	⋮
500	-2.93	0	-0.51	0.261

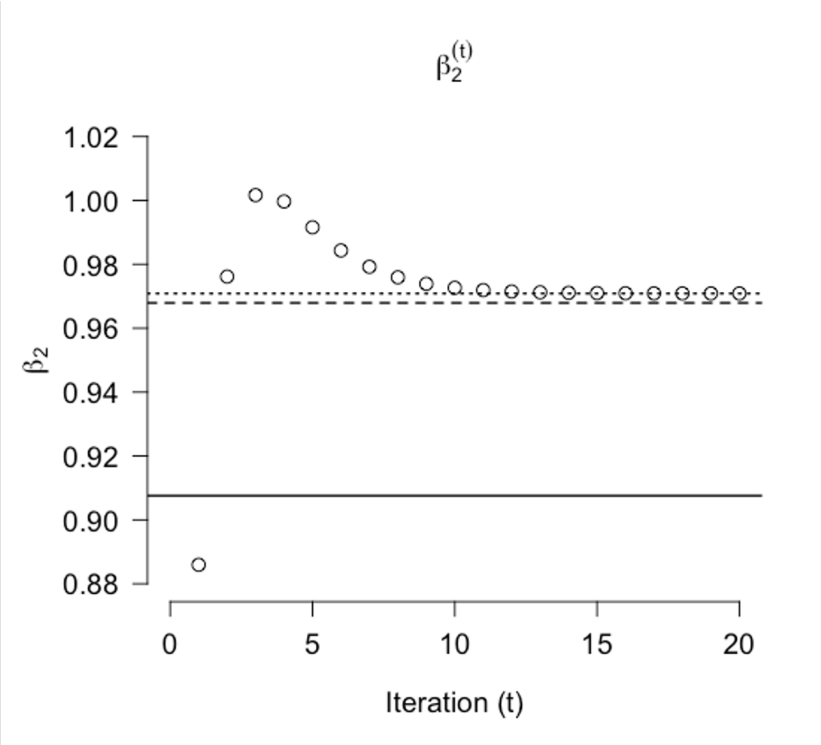
Twentieth iteration:

ID	y	x_1	$E[X_2]$	$E[X_2^2]$
1	0.29	0	0.92	1.65
2	-0.95	0	-0.12	0.82
⋮	⋮	⋮	⋮	⋮
110	3.29	1	2.84	8.87
111	-2.49	0	-1.52	2.30
⋮	⋮	⋮	⋮	⋮
500	-2.93	0	-0.51	0.261

Results: multiple linear regression with missing covariate

Parameter estimates at each iteration for a single dataset:





Standard Errors from EM

- No standard errors provided by default when using the EM algorithm
- One approach is to use the *observed information* (the negative of the second derivative of the log-likelihood) evaluated at the MLE
  - A more "peaked" log-likelihood $\rightarrow$ smaller standard errors
  - This can be a bit complicated to calculate
- Other approaches: supplemented expectation-maximization or bootstrapping

Results: multiple linear regression with missing covariate

1,000 simulated datasets of size $n = 500$, 30% missingness in X_2:

$$Y|X_1, X_2 \sim N(-1 + X_1 + X_2, 1) \quad (\beta_1 = 1,\ \beta_2 = 1)$$
$$X_2|X_1 \sim N(-1 + 2X_1, 2^2)$$

			95% CI	
Method	Coef.	Bias	Width	Coverage
Complete Data	β_1	0.00	0.25	94%
Listwise Deletion	β_1	-0.06	0.29	85%
EM Algorithm	β_1	0.00	0.32	95%
Complete Data	β_2	0.00	0.09	95%
Listwise Deletion	β_2	-0.06	0.11	48%
EM Algorithm	β_2	0.00	0.11	96%

Software Implementation in R: lavaan package

- The lavaan package in R was developed for latent variable modeling
- Default behavior is listwise deletion; however, we can apply "full information" ML estimation using the missing = 'fiml' argument (https://lavaan.ugent.be/tutorial/est.html)
- Uses a multivariate normal likelihood

Example:

```
> library(lavaan)
> sem('TotChol ~ Age + Pulse + BPSysAve', data = nha2,
+     missing = 'fiml', fixed.x = F)
```

Case Study: NHANES Results

Parameter estimates and standard errors:

$$E[\text{Total Cholesterol}|\text{Covariates}] = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Pulse} + \beta_3 \text{Syst BP}$$

	Listwise Deletion	**ML**
Intercept	2.73 (0.26)	2.68 (0.26)
Age (per 10 yrs)	0.30 (0.03)	0.31 (0.03)
Pulse (per 10 counts)	0.035 (0.022)	0.035 (0.022)
Syst BP (per 10 mmHg)	0.076 (0.020)	0.076 (0.020)

Case Study: NACC Results

Goal: investigate whether patient's APOE4 status modifies the relationship between risk factor and MMSE score

Listwise Deletion Results

	No copies	**At least one**	**P-value**
Age (per 10 yrs)	-0.36 (0.06)	-1.02 (0.08)	<0.0001
Sex (Female)	0.19 (0.09)	0.08 (0.11)	0.45
Parkinson's	-0.94 (0.31)	-2.58 (0.46)	0.003
Depression	-0.62 (0.13)	0.25 (0.15)	<0.0001
Stroke	-1.46 (0.20)	-1.72 (0.26)	0.43
Clinical AD	-5.07 (0.10)	-5.94 (0.11)	<0.0001

Case Study: NACC Results

ML Results

	No copies	At least one	P-value
Age (per 10 yrs)	-0.37 (0.06)	-1.20 (0.08)	<0.0001
Sex (Female)	0.22 (0.09)	-0.02 (0.11)	0.11
Parkinson's	-0.89 (0.28)	-2.87 (0.45)	0.0005
Depression	-0.70 (0.12)	0.25 (0.15)	<0.0001
Stroke	-1.51 (0.20)	-1.90 (0.26)	0.28
Clinical AD	-5.02 (0.10)	-6.41 (0.11)	<0.0001

Auxiliary Variables

- **Auxiliary variables**: variables thought to be associated with missingness or with a variable that has missingness
- Not usually of scientific interest and thus would not want to be included in the analysis model (doing this would change interpretations)
- By including them in the procedure, can improve efficiency and reduce bias (if needed for validity of MAR assumption)
- One approach is the **saturated correlates model** (Graham 2003)
  - See `sem.auxiliary()` in the `semTools` package

Longitudinal Data

In longitudinal studies, missing responses can be common (e.g., due to dropout)

- An assumption (MAR) commonly made is that the probability of dropout depends on prior response (and covariates)
- Can obtain valid inference under MAR using maximum likelihood estimation of mixed models

ID	X_1	X_2	Y_1	Y_2	Y_3
1	5.20	1	1.23	2.31	1.87
2	4.62	0	0.73	0.81	0.82
3	6.45	1	0.52		
4	2.88	1	1.02	1.45	1.68
5	3.56	0	0.92	0.86	
6	4.21	0	1.45	1.21	

Summary

- Define a model for the variable(s) with missing values to obtain the *complete-data* likelihood
- This involves integrating over the distribution of the missing values and then maximizing this *observed-data* likelihood
- This can be challenging, a common approach used is the **EM algorithm**:
  - **E-step**: compute the expectation of the *complete-data* log-likelihood conditional on the observed data and current parameter estimates
  - **M-step**: update the parameter estimates by maximizing this conditional expected log-likelihood
- Assumes sample size is large enough for consistency and approximate normality
- Assumes a parametric model for the data

End of Day 1

Overview

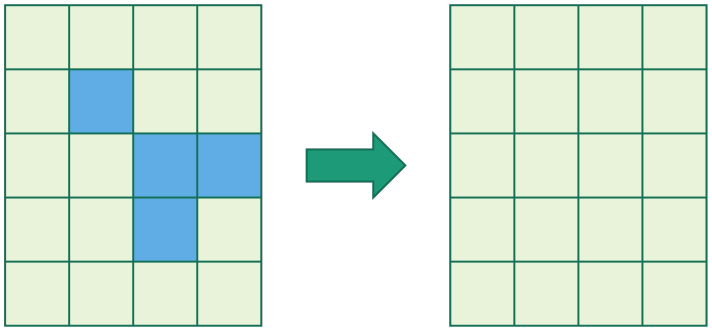
The Big Picture

- Maximum likelihood is a natural way to handle missing data under the assumption of MAR, but can be challenging to implement in some situations
- Multiple imputation (MI) has similar properties as that of ML, but can be easier to operationalize

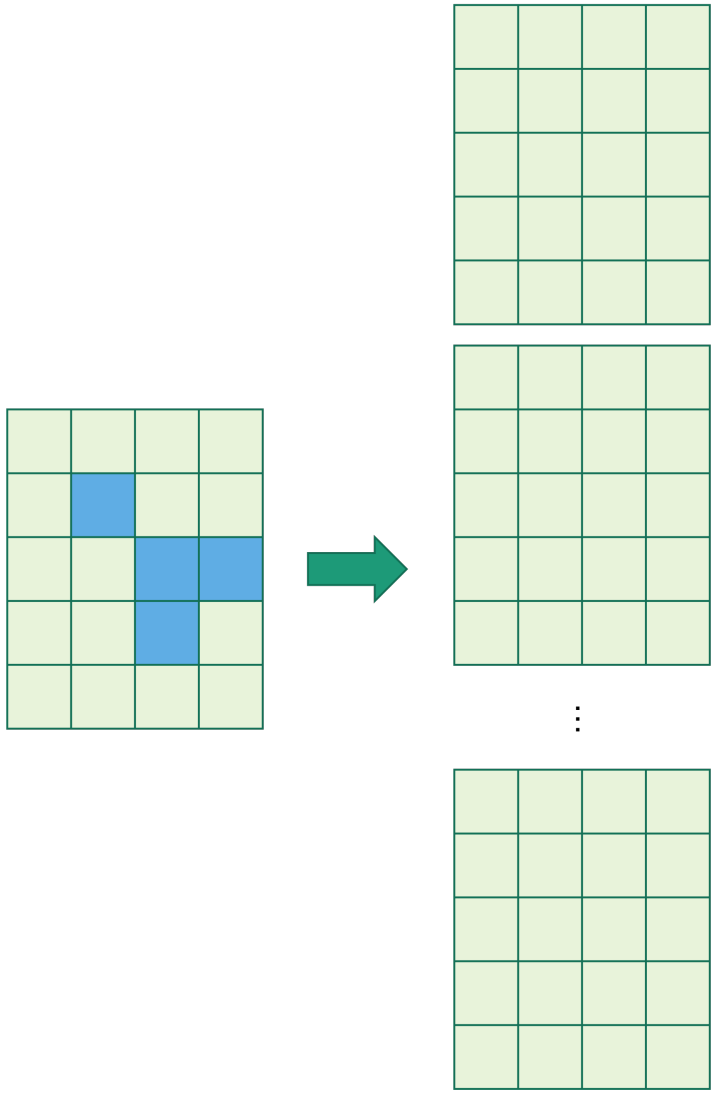
  - Imputing or "filling in" missing values
  - Information about the values can be obtained from other observed variables
  - In **multiple imputation** we create many imputed datasets

Single vs Multiple Imputation

Single imputation:



Multiple imputation:

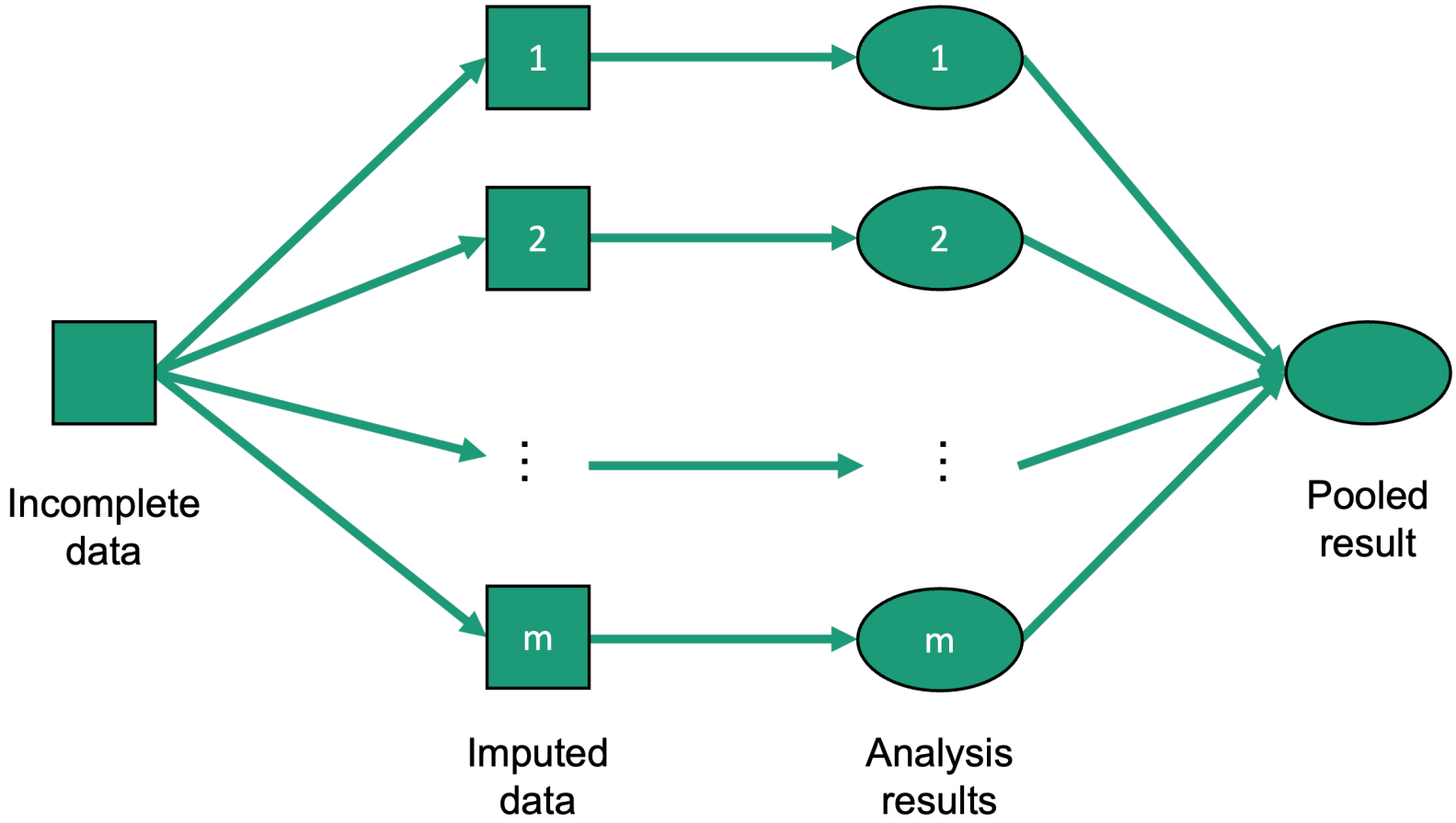


Single Imputation: Reminder

Method	Coefficient Estimates	Standard Error
Mean	Biased	Too small
Regression	Valid under MAR	Too small
Stochastic	Valid under MAR	Too small

MI can help deal with too small standard errors by helping us account for the uncertainty of the imputations

Multiple Imputation



Adapted from `https://stefvanbuuren.name/fimd/sec-nutshell.html`

Multiple Imputation: Step-by-Step

Step 0 Start with incomplete data set

Step 1 Plausible values drawn from distribution modeled for missing variables – do this m times to create m imputed datasets

X_1	X_2	X_3	Z
4	0	65	10
5	0	60	NA
4	1	70	12
7	1	72	16
3	1	60	NA

Incomplete data

X_1	X_2	X_3	Z
4	0	65	10
5	0	60	10
4	1	70	12
7	1	72	16
3	1	60	9

X_1	X_2	X_3	Z
4	0	65	10
5	0	60	12
4	1	70	12
7	1	72	16
3	1	60	10

⋮

X_1	X_2	X_3	Z
4	0	65	10
5	0	60	10
4	1	70	12
7	1	72	16
3	1	60	10

Multiple Imputation: Step 1 Detail

- In this step, we need to specify an imputation model
- We want one that produces **proper imputations**
- The goal is to draw values of the missing variable from $f(\mathbf{z}^{(mis)}|\mathbf{z}^{(obs)}, \mathbf{x})$ (called the *posterior predictive distribution*)
  - We are assuming that missingness is MAR so $f(\mathbf{z}^{(mis)}|\mathbf{z}^{(obs)}, \mathbf{x}, \mathbf{r}) = f(\mathbf{z}^{(mis)}|\mathbf{z}^{(obs)}, \mathbf{x})$
- Proper imputation should account for uncertainty about the parameters in the model
  - For example, in the stochastic regression imputation approach (slide 73), the coefficient values in the regression used for generating imputations were fixed at their estimates; however, these are sample estimates themselves
  - Should base imputations on random draws from the *Bayesian posterior distribution* of the parameters

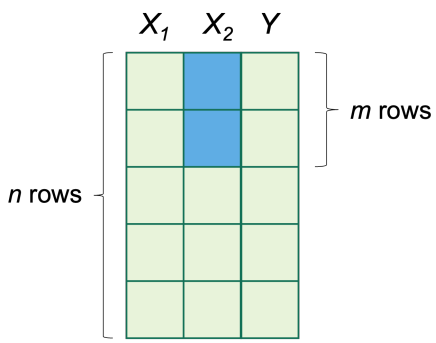
Multiple Imputation: Step 1 Detail

$$Y = \beta_0 + \beta_1 X + \epsilon; \quad \epsilon \sim N(0, \sigma^2)$$

- Approximately 30% missingness in Y, depends on X
- $n = 50$
- $m = 10$
- 1,000 replications

	β_0			β_1		
Approach	**Bias**	**Width**	**Coverage**	**Bias**	**Width**	**Coverage**
SI - stochastic reg	0.00	0.44	84%	0.00	0.45	82%
MI - stochastic reg	0.00	0.51	92%	0.00	0.51	91%
MI - proper	0.00	0.58	94%	0.00	0.58	94%
Listwise deletion	0.00	0.56	94%	0.00	0.56	94%

Multiple Imputation: Step 1 Example



Suppose we are interested in estimating β_0, β_1, and β_2:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2)$$

where X_{2i} is missing for m observations.

How to create imputations under a normal linear model?

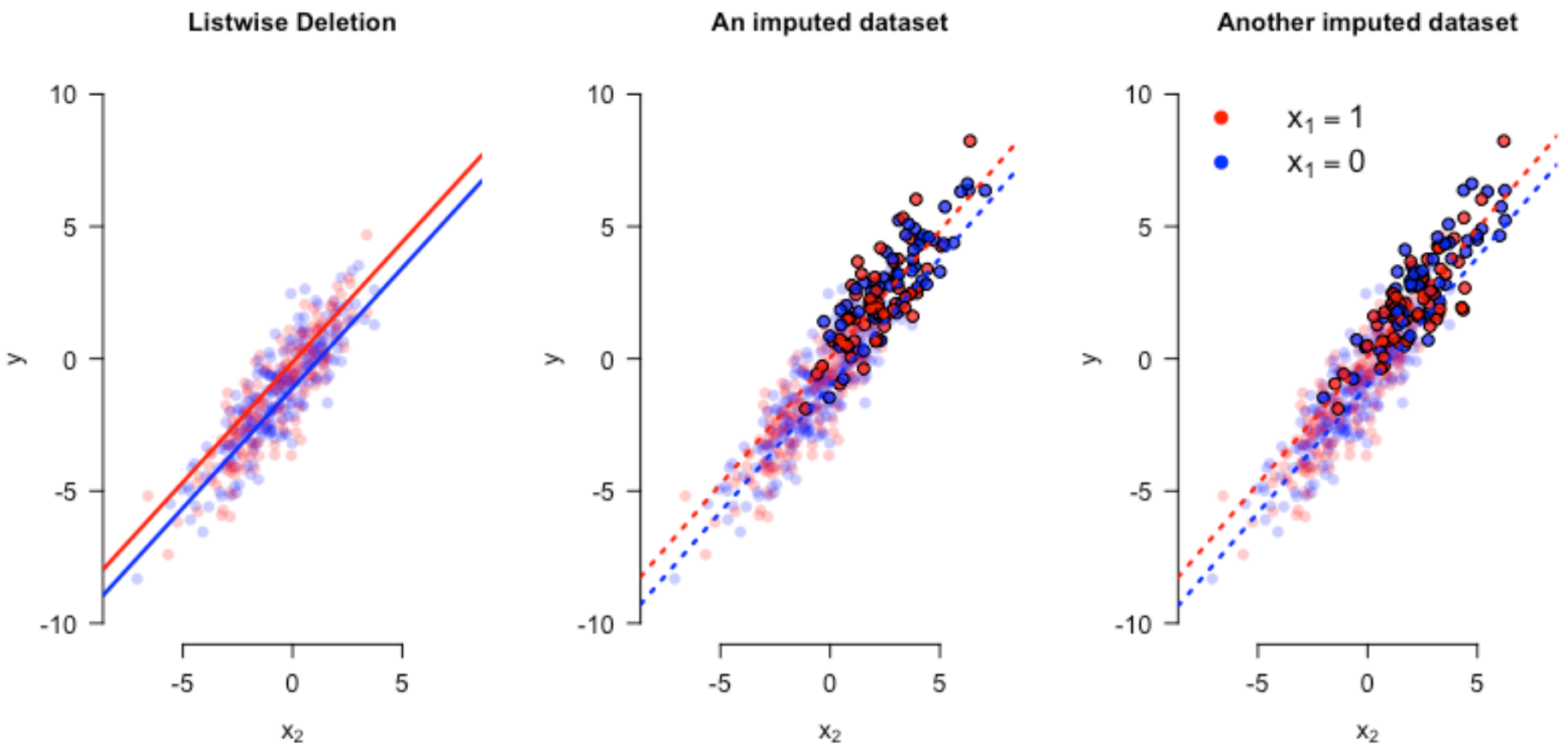
1. Fit the linear model to the $n - m$ rows with full data:

$$X_{2i} = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 Y_i + \delta_i; \quad \delta_i \sim N(0, \tau^2)$$

2. Randomly draw $\dot{\gamma}_0, \dot{\gamma}_1, \dot{\gamma}_2, \dot{\tau}^2$ from their *posterior distribution*
3. For the m rows with missingness create imputations:

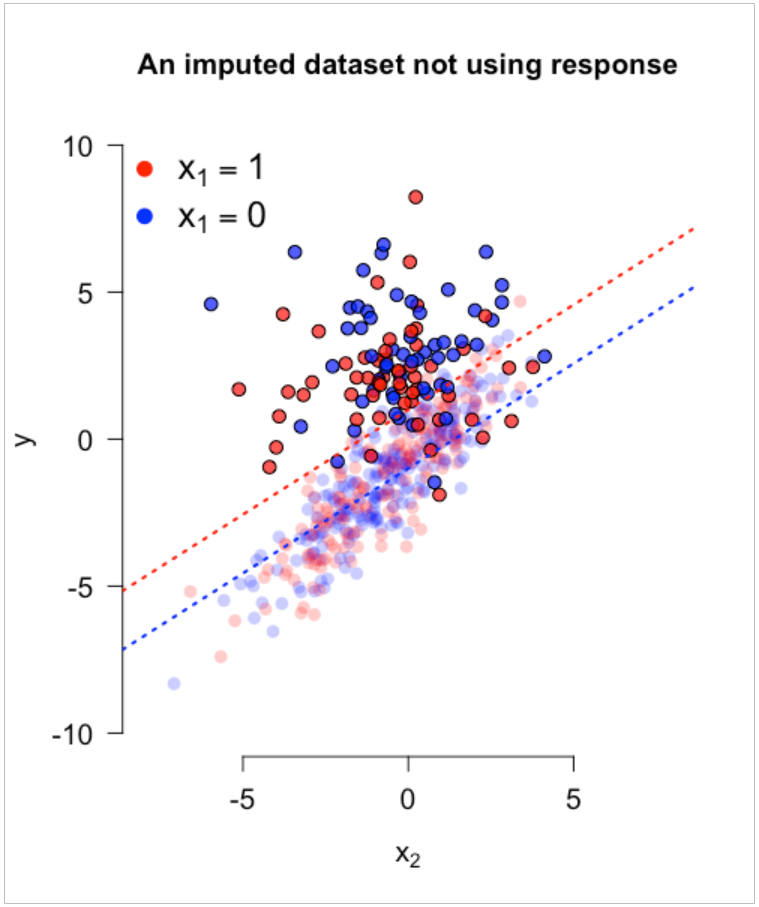
$$\dot{X}_{2i} = \dot{\gamma}_0 + \dot{\gamma}_1 X_{1i} + \dot{\gamma}_2 Y_i + \dot{\delta}_i; \quad \dot{\delta}_i \sim N(0, \dot{\tau}^2)$$

Multiple Imputation: Step 1 Example



Multiple Imputation: Step 1 Example

Suppose we didn't include the response when creating the imputations:



Multiple Imputation: Step 1 Detail

Some common imputation models (and the `method` argument to use in the `R` package `mice`):

- Normal linear model: `method = 'norm'` and `method = 'norm.boot'`
- Non-normal distributions:
  - Predictive mean matching: `method = 'pmm'`
  - Others...
  - Work has shown that for inference of regression coefficients, normal imputations can be robust
  - Could transform, but sometimes that can make things worse
- Logistic regression model: `method = 'logreg'`
  - Multinomial logit model: `method = 'polyreg'`
  - Orderd logit model: `method = 'polr'`

Multiple Imputation: Step 1 Detail

Practical considerations: what variables to include in the imputation model?

- Variables intended to be used in analysis model (including the outcome variable)
- Auxiliary variables: variables highly correlated with those that have missingness or with probability of missingness
  - MI can generally handle a larger number of auxiliary variables so the typical advice is to include many

Multiple Imputation: Step 1 Detail

Validity of imputation model?

- It can be fairly robust to misspecification because the imputation model is only applied to the observations with missingness (and not the entire data set)
- Trouble can arise if imputation model is too simple (e.g., analysis model includes interactions or polynomials and imputation model does not)
- "Superefficiency" if imputer has additional knowledge that can inform the imputation model

Multiple Imputation: Step 1 Detail

What happens when we have missingness in more than one variable?

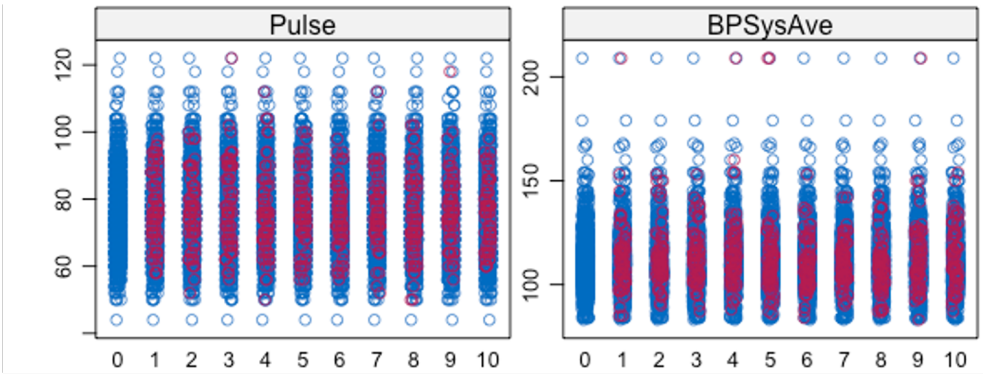
- **Monotone missingness**: imputations drawn from sequence of univariate regressions
  - No need for more than 1 iteration
- **Joint modeling**: imputations drawn from multivariate model
- **Fully conditional specification** (aka **chained equations**): imputations drawn from sequence of univariate regressions and iterated

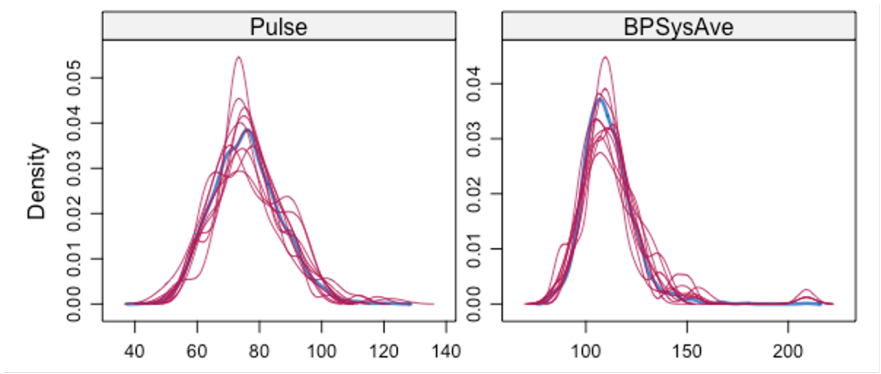
Case Study: NHANES

Regression models used:

- $E[\text{Pulse}|\text{TotChol}, \text{Age}] = \gamma_0 + \gamma_1 \text{TotChol} + \gamma_2 \text{Age}$
- $E[\text{BPSysAve}|\text{TotChol}, \text{Age}, \text{Pulse}] = \gamma_0^\star + \gamma_1^\star \text{TotChol} + \gamma_2^\star \text{Age} + \gamma_3^\star \text{Pulse}$

```
> step1 <- mice(nha2 %>% select(TotChol, Age, Pulse, BPSysAve),
+               maxit = 1, m = 10, visitSequence = "monotone",
+               method = "pmm", seed = 6, print = F)
```



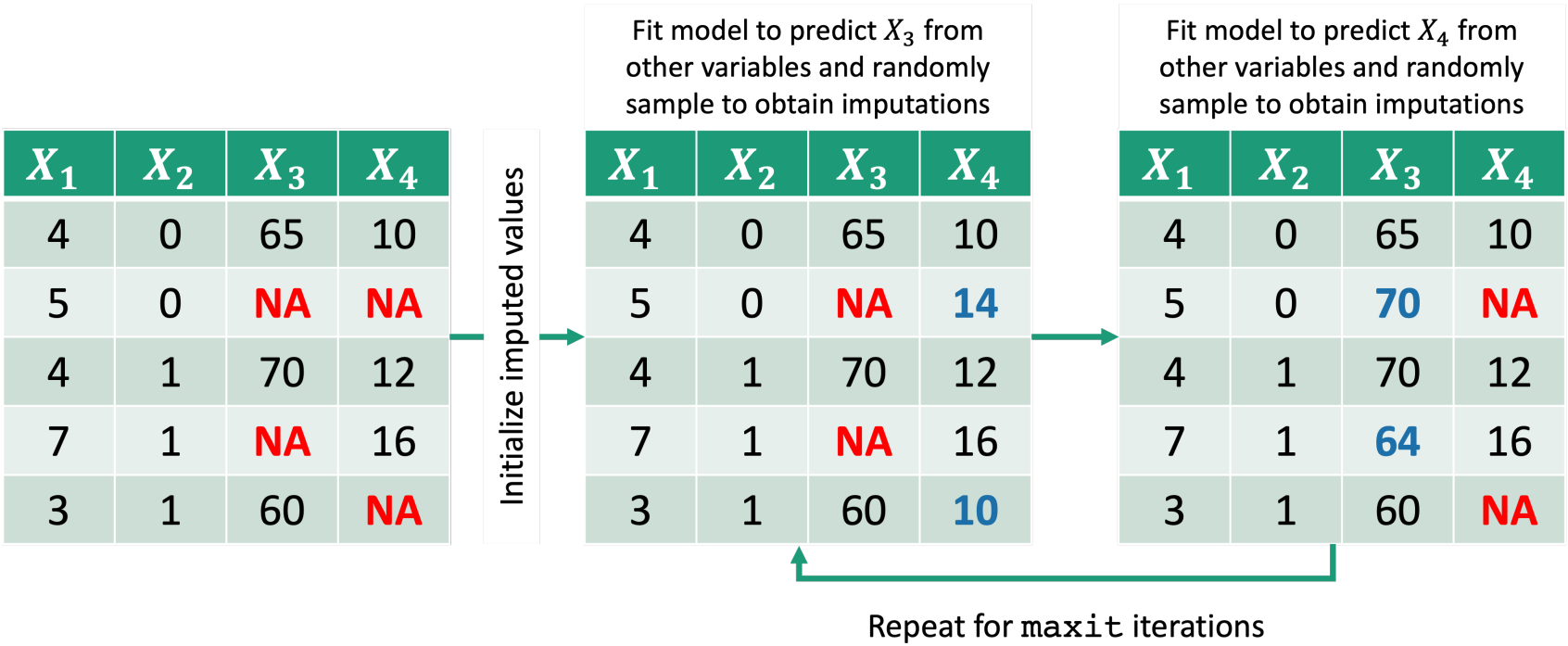


Joint Modeling

- Specify a multivariate distribution for the data
  - A common choice is a multivariate normal distribution
  - Note: it can be challenging to specify a joint distribution when in the setting of a mixture of categorical and continuous data
- An approach known as **data augmentation** can be used to obtain imputed values using an iterative approach:
  - (I step) Randomly sample missing data from it's distribution conditional on observed data and current values of imputation model parameters
  - (P step) Randomly sample imputation model parameters given current random draw of missing data
- The `jomo` package in `R` was developed specifically for joint modeling

Fully Conditional Specification

- Basic idea: apply procedure used for monotone missingness (impute one by one), but iterate
- Lacks theoretical underpinnings; however, appears to do well in simulation
- Flexible since not constrained to certain multivariate distributions



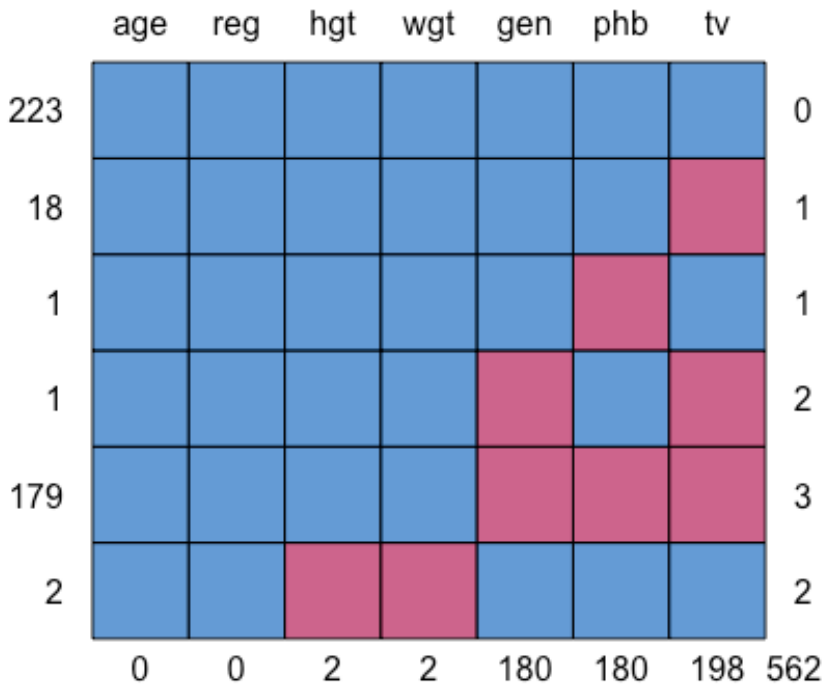
X_1	X_2	X_3	X_4
4	0	65	10
5	0	NA	NA
4	1	70	12
7	1	NA	16
3	1	60	NA

X_1	X_2	X_3	X_4
4	0	65	10
5	0	NA	14
4	1	70	12
7	1	NA	16
3	1	60	10

X_1	X_2	X_3	X_4
4	0	65	10
5	0	70	NA
4	1	70	12
7	1	64	16
3	1	60	NA

Example

- Subsample of 424 boys between ages 8-21; available from the `boys` dataset in `mice`
- Interest was is in development of secondary pubertal characteristics
  - `gen`: five ordered stages of genital development
  - `phb`: six ordered stages of pubic hair development
  - `tv`: testicular volume
- Missingness related to age



Proportion missing `gen` by age:

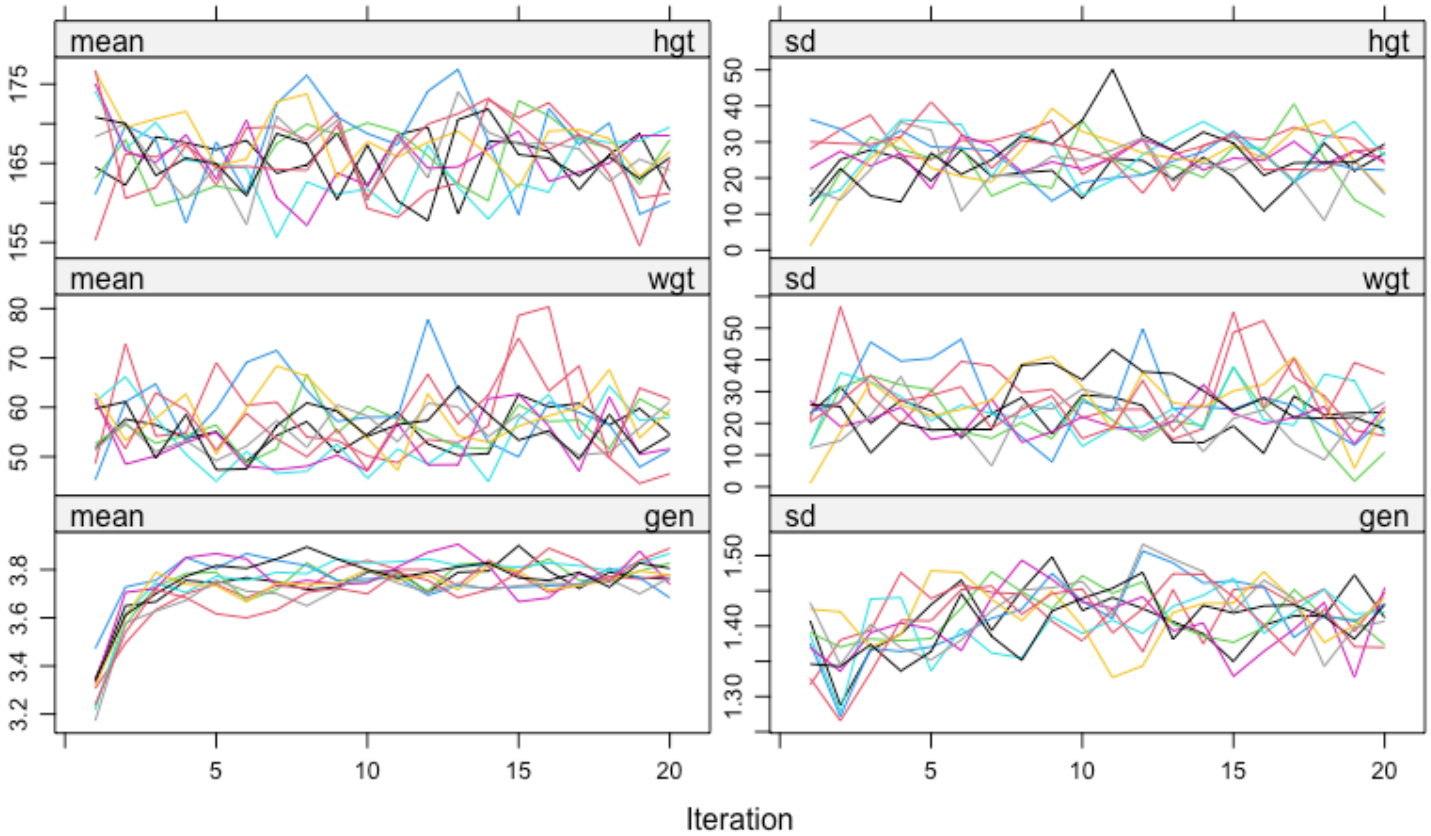
8–11	11–14	14–17	17–21
24%	33%	46%	58%

Example

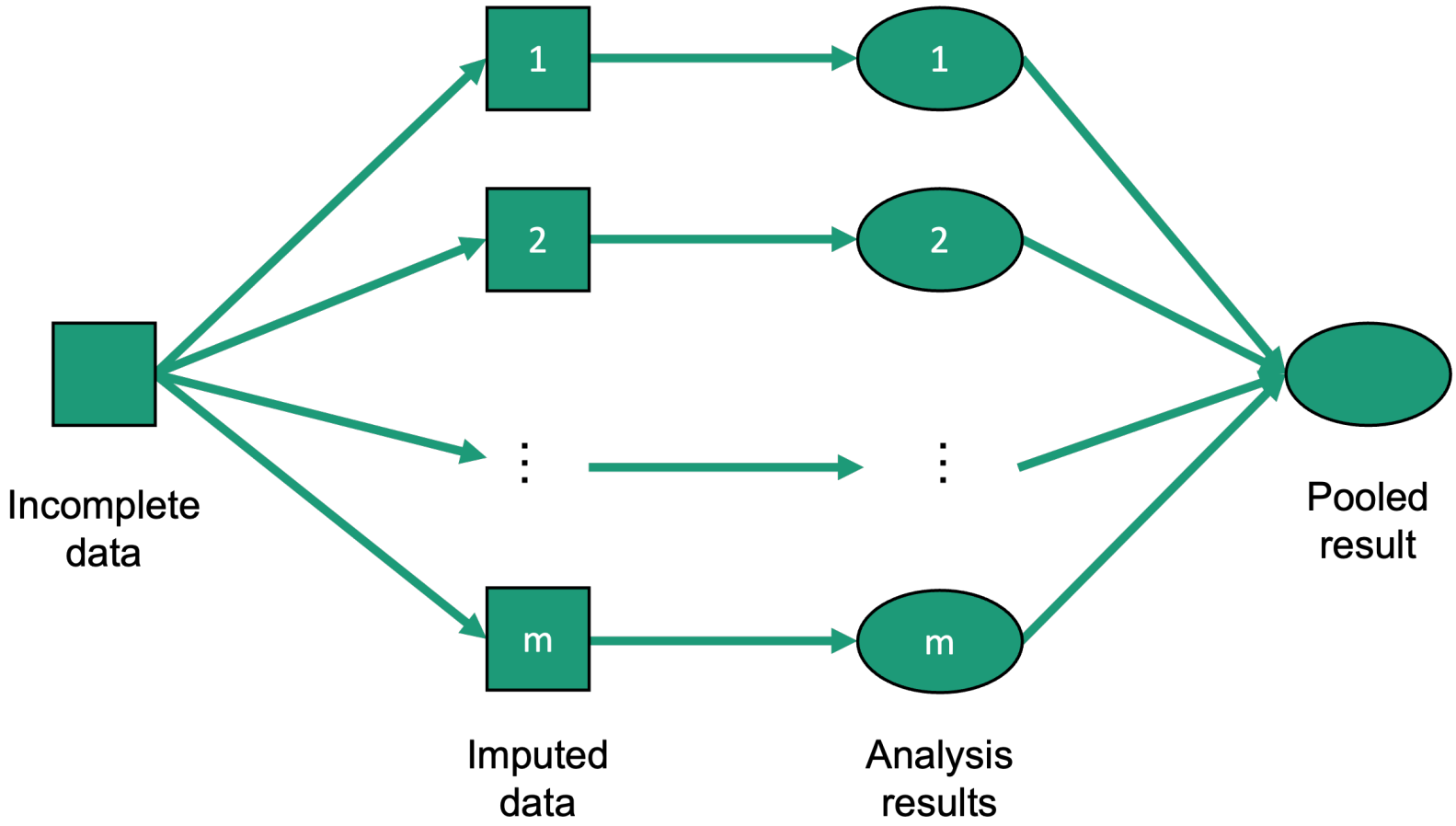
```
> fcs <- mice(boys_subset, seed = 6, m = 10, maxit = 20,
print = FALSE)
> fcs$predictorMatrix
age hgt wgt gen phb tv reg
age   0   1   1   1   1  1   1
hgt   1   0   1   1   1  1   1
wgt   1   1   0   1   1  1   1
gen   1   1   1   0   1  1   1
phb   1   1   1   1   0  1   1
tv    1   1   1   1   1  0   1
reg   1   1   1   1   1  1   0
> fcs$method
age     hgt     wgt     gen     phb      tv     reg
""   "pmm"   "pmm" "polr" "polr"   "pmm"      ""
```

Example

Plots of parameters vs iteration numbers can be useful for diagnosing convergence issues:



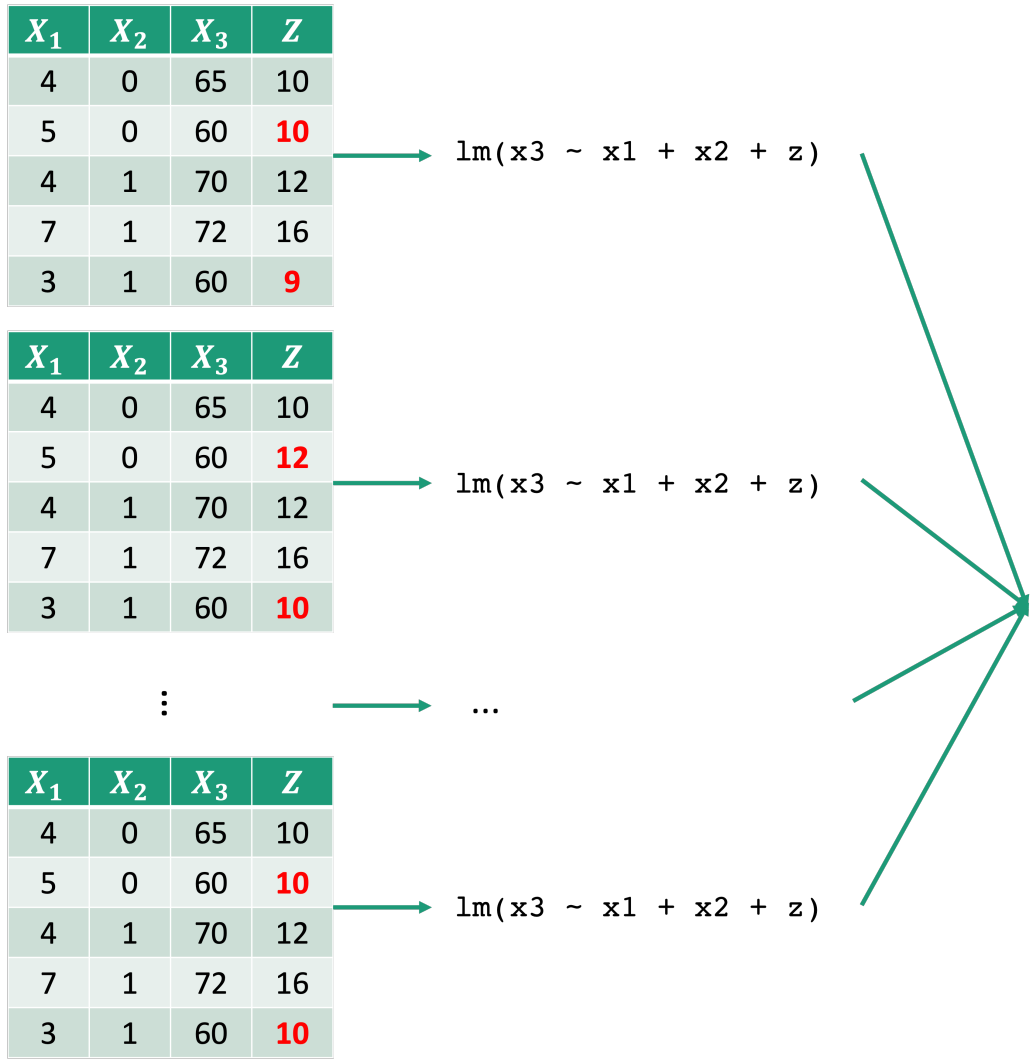
Multiple Imputation



Multiple Imputation: Step-by-Step

Step 2 Analyze each of the m datasets separately

Step 3 Pool results of m analyses



Case Study: NHANES

```
> step2 <- with(step1, lm(TotChol ~ Age + Pulse + BPSysAve))
```

Dataset	β_0	β_1	β_2	β_3
1	2.83 (0.25)	0.032 (0.0031)	0.0028 (0.0021)	0.0066 (0.0019)
2	2.72 (0.27)	0.031 (0.0031)	0.0032 (0.0021)	0.0074 (0.0019)
3	2.70 (0.26)	0.031 (0.0031)	0.0027 (0.0021)	0.0080 (0.0019)
⋮	⋮	⋮	⋮	⋮
10	2.64 (0.26)	0.031 (0.0031)	0.0039 (0.0021)	0.0077 (0.0019)

Multiple Imputation: Step 3 Detail

- In the final step, we **pool** the results using **Rubin's rules**
- Pooled parameter estimate = average of the estimates from the m analyses

$$\hat{\theta} = \frac{1}{m}\sum_{d=1}^{m}\hat{\theta}_d$$

- Standard errors follows same logic, but more complicated
  - Two sources of sampling variability: (1) "usual" variability (what would have resulted if data were complete) and (2) variability due to missing data
  - The total sampling variance is:

$$\underbrace{\frac{1}{m}\sum_{d=1}^{m}\text{SE}_d^2}_{\text{Within-imputation}} + \left(1+\frac{1}{m}\right)\underbrace{\frac{1}{m-1}\sum_{d=1}^{m}(\hat{\theta}_d-\hat{\theta})^2}_{\text{Between-imputation}}$$

Case Study: NHANES

```
> step3 <- pool(step2)
> step3
Class: mipo    m = 10
       term  m estimate    ubar       b       t dfcom   df   riv lambda    fmi
(Intercept) 10   2.7284 6.6e-02 6.4e-03 7.3e-02  1499  566 0.106  0.096 0.0993
Age         10   0.0315 9.9e-06 4.5e-08 9.9e-06  1499 1483 0.005  0.005 0.0064
Pulse       10   0.0033 4.4e-06 3.4e-07 4.8e-06  1499  707 0.085  0.079 0.0813
BPSysAve    10   0.0073 3.5e-06 2.7e-07 3.8e-06  1499  722 0.084  0.077 0.0796
```

- `estimate`: $\hat{\theta}$
- `ubar`, `b`, and `t`: components of the variance
- `df` corrected degrees of freedom
- `riv`: relative increase in variance due to nonresponse
- `lambda`: proportion of variance due to nonresponse
- `fmi`: fraction of missing information due to nonresponse
  - 0.2: "modest"
  - 0.3: "moderately large"
  - 0.5: "high" – inference strongly depends on how missingness was addressed

Case Study: NHANES Results

$$E[\text{Total Cholesterol}|\text{Covariates}] = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Pulse} + \beta_3 \text{Syst BP}$$

	LD	**ML**	**MI**
Intercept	2.73 (0.26)	2.68 (0.26)	2.73 (0.27)
Age (per 10 yrs)	0.30 (0.03)	0.31 (0.03)	0.32 (0.03)
Pulse (per 10 counts)	0.035 (0.022)	0.035 (0.022)	0.033 (0.022)
Syst BP (per 10 mmHg)	0.076 (0.020)	0.076 (0.020)	0.073 (0.020)

How Large Should m Be?

- It turns out that MI can work pretty well with small m; in many practical applications, $m = 20$ is reasonable
- Use larger m if the *missing information rate* is high
- Do not need large m for precise estimates of regression coefficients, larger m needed for other measures (e.g. p-values)
- Nice summary:
  `https://stefvanbuuren.name/fimd/sec-howmany.html`

Simulation

Based on 1,000 simulated datasets of size n $= 500$, $\rho(X_1, X_2) = 0.5$, $m = 10$, 40% missingness:

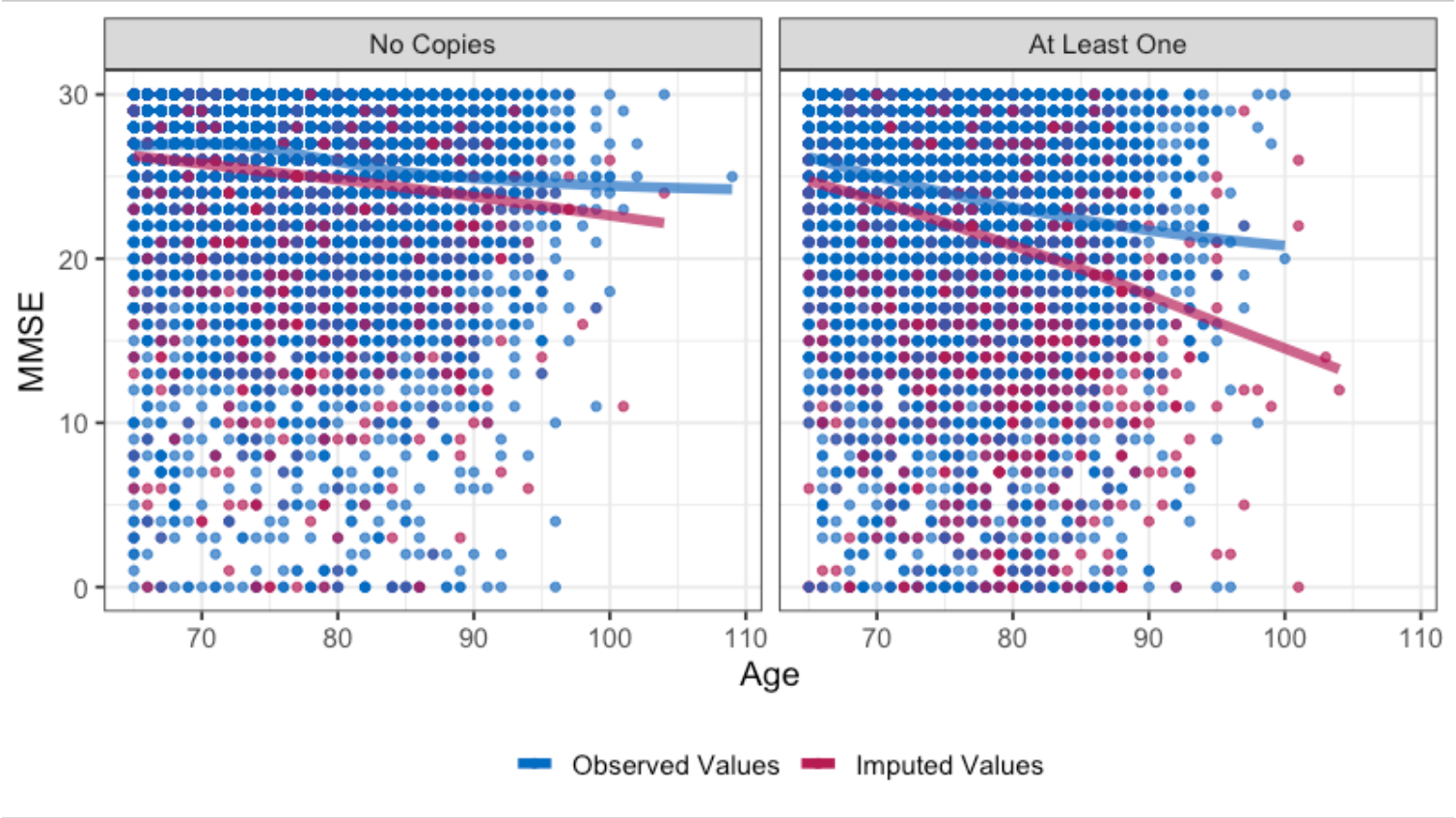
Setting	Coef.	Bias	SD($\hat{\beta}$)	95% CI Coverage
Complete Data	β_1	0	0.044	95%
MCAR	β_1	0	0.059	94%
Missing $Y \mid X_2$	β_1	0	0.057	95%
Missing $X_2 \mid Y$	β_1	0	0.053	95%
Complete Data	β_2	0	0.041	97%
MCAR	β_2	0	0.056	96%
Missing $Y \mid X_2$	β_2	0	0.066	95%
Missing $X_2 \mid Y$	β_2	0	0.056	96%

Case Study: NACC Results

MI Results ($m = 20$ imputations)

	No copies	**At least one**	**P-value**
Age (per 10 yrs)	-0.38 (0.06)	-1.20 (0.08)	<0.0001
Sex (Female)	0.22 (0.09)	-0.02 (0.11)	0.11
Parkinson's	-0.89 (0.28)	-2.87 (0.44)	0.0003
Depression	-0.71 (0.13)	0.25 (0.15)	<0.0001
Stroke	-1.57 (0.20)	-1.80 (0.26)	0.51
Clinical AD	-5.03 (0.11)	-6.42 (0.11)	<0.0001

Case Study: NACC Dataset



Longitudinal Data

- With dropout, the missingness pattern is monotone. Thus, in "wide" format can proceed by using a series of regression models where "current" response is regressed on previous responses plus covariates
- MI can also be done in "long" format using multilevel imputation (to account for correlation)

ID	X_1	X_2	Y_1	Y_2	Y_3
1	5.20	1	1.23	2.31	1.87
2	4.62	0	0.73	0.81	0.82
3	6.45	1	0.52		
4	2.88	1	1.02	1.45	1.68
5	3.56	0	0.92	0.86	
6	4.21	0	1.45	1.21	

Summary

- Can think of MI as extending likelihood methods where an extra step is introduced (drawing imputed values)
- MI performance similar to likelihood method that makes similar assumptions
- Can be easier to calculate standard errors; can inspect the imputations to help evaluate modeling assumptions
- Also relies on large sample approximations
  - Some work has shown that approximations may work better for MI than ML with small to moderate sample sizes

Summary

- Valid under MAR mechanism
- Three major steps:
  1. Estimate a *posterior predictive distribution* for the missing values and fill in (impute) missing values with draws from the predictive distribution
  2. Repeat $m > 1$ times and apply the final analysis model to the m datasets to get m estimates
  3. Combine the results to get an overall estimate
- Benefit of MI:
  - Allows variables not included in the final analysis model to be included in the imputation model
  - Can be extended to MNAR settings

Overview

Intuition

Suppose we have the following data from a random sample of a population (half exposed, half unexposed):

Group	Exposed	Unexposed
Response	0 0 0 0	1 1 1 1

The average response is 0.5.

What if only one individual in the exposed group and 3 in the unexposed group responded?

Group	Exposed	Unexposed
Response	0 ? ? ?	1 1 1 ?

If we calculate the average response of those observed we would get 0.75 $\rightarrow$ bias!

Intuition

Suppose we have the following data from a random sample of a population (half exposed, half unexposed):

Group	Exposed	Unexposed
Response	0 0 0 0	1 1 1 1

The average response is 0.5.

What if only one individual in the exposed group and 3 in the unexposed group responded?

Group	Exposed	Unexposed
Response	0 ? ? ?	1 1 1 ?

If we calculate the average response of those observed we would get 0.75 $\rightarrow$ bias!

Intuition

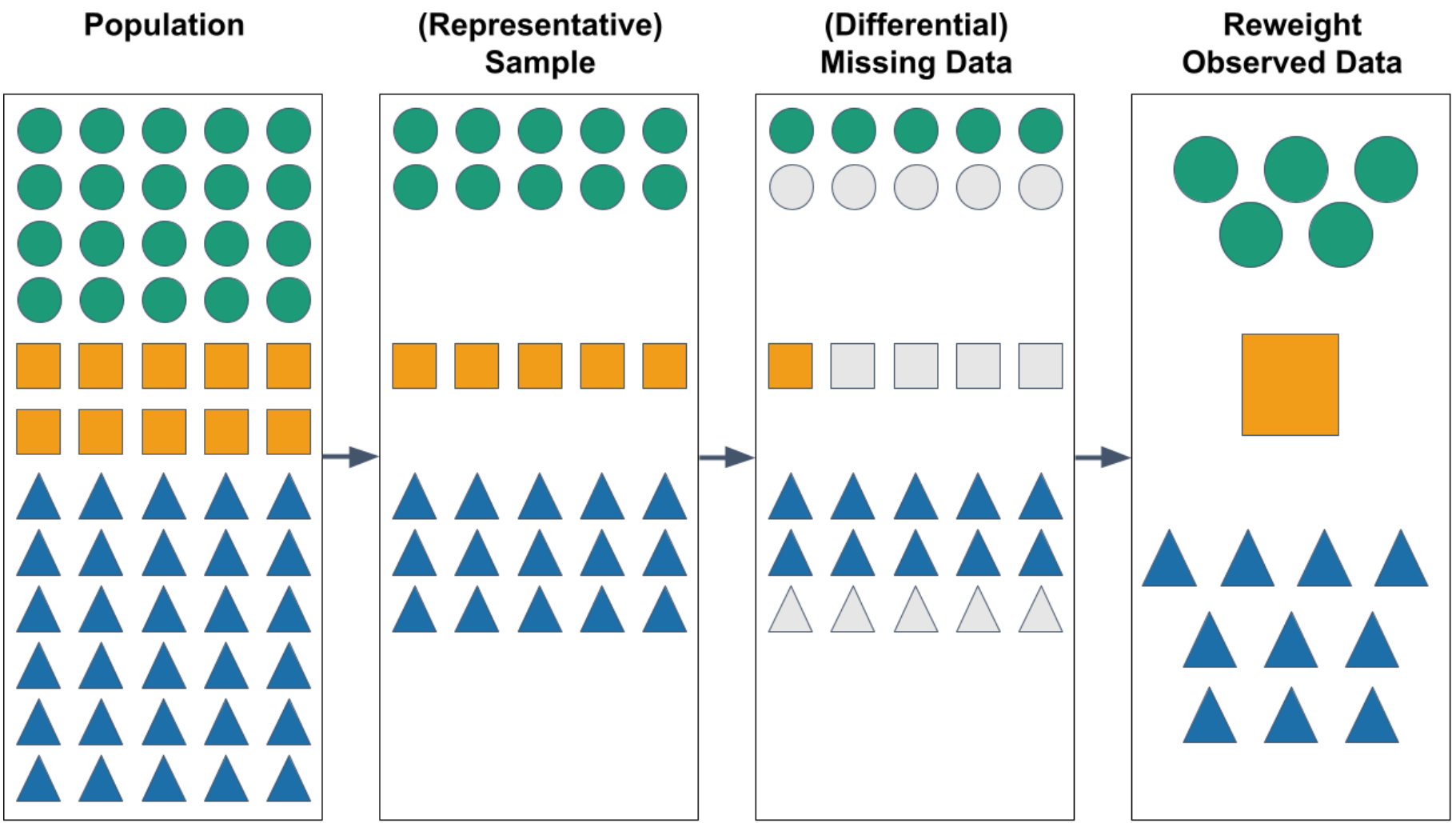
Group	Exposed	Unexposed
Response	0 ? ? ?	1 1 1 ?

Instead, we can calculate a weighted average, weighting by the inverse of the probability of being observed:

$$\frac{0 \times (4/1) + (1 + 1 + 1) \times (4/3)}{(4/1) + (4/3) + (4/3) + (4/3)} = 0.5$$

Inverse probability weighting (IPW) has eliminated the bias!

Intuition


Population
(Representative) Sample
(Differential) Missing Data
Reweight Observed Data

Imputation vs. Weighting

There are two key differences between weighting and imputation:

1. Weighting is **semiparametric:** doesn't require specifying full likelihood

2. Weighting approaches model the missing data mechanism rather than integrating over the missing data distribution

Reminder: Notation

- Variables:
  - X: predictors/covariates of interest
  - V: auxiliary variables that may affect $P(R = 1)$, but are not interesting for analysis
  - Y: outcome/response, which may have missingness (i.e., $Z = Y$ in this case)
  - R: indicator for missingness ($R = 1$ if Y is missing)
  - For simplicity, define $C = 1 - R$ (so $C = 1$ if Y is observed)

- Parameters:
  - ϕ: parameters that govern the missing data process
  - θ: parameters of scientific interest

Intuition

- Suppose we know $P(C = 1|Y, V, \phi) = P(C = 1|V, \phi)$
  - i.e., probability of observing Y is known and depends only on observed variables V
- Call $P(Y \text{ observed}) = P(C = 1|V, \phi) = \pi(V)$
- If one of our observed cases (Y such that $C = 1$) was particularly unlikely to be observed (low $\pi(V)$), we'll upweight it to represent all the similar cases we *didn't* observe
  - In fact, we'll weight it by $\frac{1}{\pi(V)}$
  - This is called **inverse probability weighting**

Set-Up: Estimation of Mean from Full Data

- Simplest case: consider estimating a mean, $\mu = E(Y)$
- With no missing data, we'd use the sample mean,

$$\hat{\mu}^{full} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

- Note that $\hat{\mu}^{full}$ is the solution to the **estimating equation**

$$\sum_{i=1}^{N}(Y_i - \mu) = 0$$

Set-Up: Complete Case Analysis

- Because $\pi(V)$ depends on observed V, missingness is MAR
- We've already seen that the complete case estimator,

$$\hat{\mu}^{CC} = \frac{\sum_{i=1}^{N} C_i Y_i}{\sum_{i=1}^{N} C_i},$$

which solves the estimating equation

$$\sum_{i=1}^{N} C_i (Y_i - \mu) = 0,$$

is typically not consistent under MAR (with auxiliary variables V)

Inconsistency of CC Estimator

A sketch of the proof that $\hat{\mu}^{CC}$ is generally not consistent:

$$\hat{\mu}^{CC} = \frac{\sum_{i=1}^{N} C_i Y_i}{\sum_{i=1}^{N} C_i} \xrightarrow{p} \frac{E(CY)}{E(C)} \qquad \text{by LLN}$$

$$= \frac{E\{Y\pi(V)\}}{E\{\pi(V)\}} \qquad \text{using iterated expectations}$$

$$\neq E(Y) = \mu \qquad \text{if } Y, V \text{ not independent}$$

(See Appendix for details.)

Inverse Probability Weighting (IPW)

- We can derive a consistent estimator (in fact, an unbiased one!) by weighting the complete case estimating equation:

$$\sum_{i=1}^{N} \frac{C_i}{\pi(V_i)}(Y_i - \mu) = 0$$

  - Intuition: $E(C_i|Y_i, V_i) = \pi(V_i)$, so these terms now cancel out in the expectation (see Appendix for details)

- Estimator that solves this estimating equation is

$$\hat{\mu}^{ipw} = \left\{ \sum_{i=1}^{N} \frac{C_i}{\pi(V_i)} \right\}^{-1} \sum_{i=1}^{N} \frac{C_i Y_i}{\pi(V_i)}$$

- Huber-White "robust" sandwich estimator is often used for standard error estimation, but is typically conservative
  - Recommended alternative: bootstrap or jackknife

What if Missingness Mechanism is Unknown?

- So far, we have treated $\pi(V_i)$ as known
  - E.g., missingness by design (rotating panel, subset of participants chosen for expensive measure, etc.)

- In practice, this is often not true

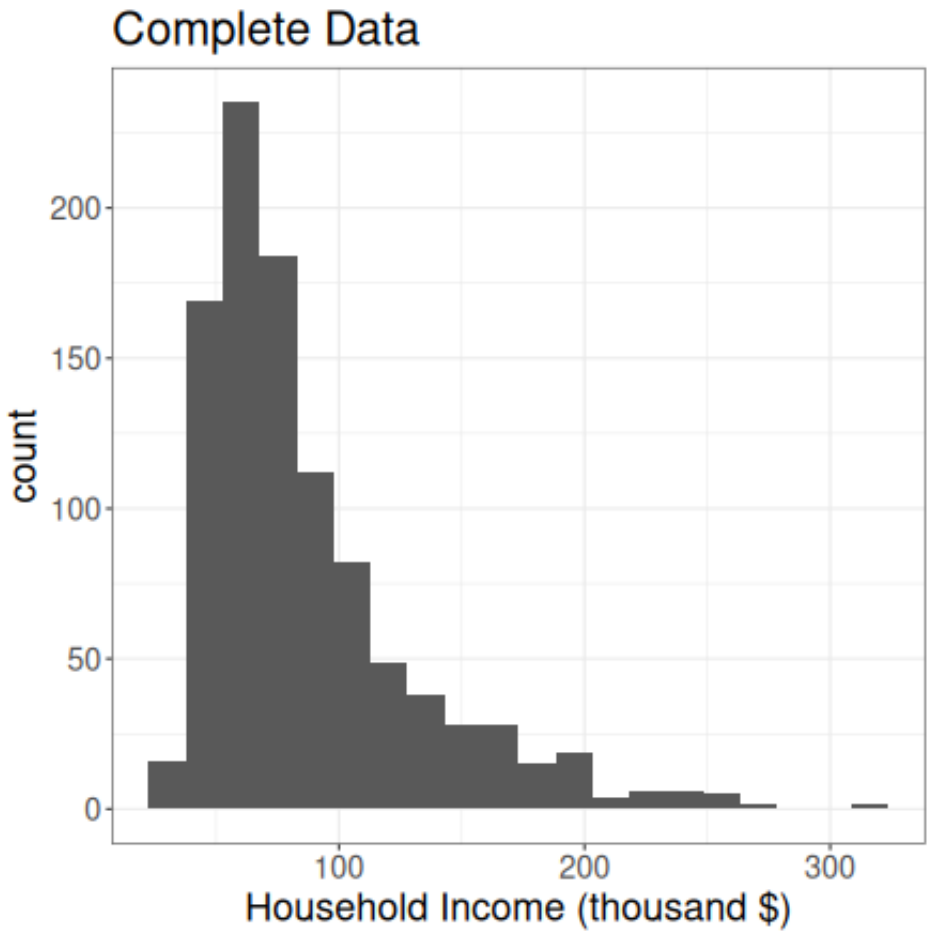
- We instead model $\pi(V_i) = P(C_i|V_i;\phi)$ parametrically using, e.g., logistic regression
  - If missingness model is correctly specified, using estimated weights still yields consistent estimator:

$$\hat{\mu}^{ipw} = \left\{ \sum_{i=1}^{N} \frac{C_i}{\pi(V_i;\hat{\phi})} \right\}^{-1} \sum_{i=1}^{N} \frac{C_i Y_i}{\pi(V_i;\hat{\phi})}$$

  - If missingness model is misspecified, IPW estimator may be inconsistent

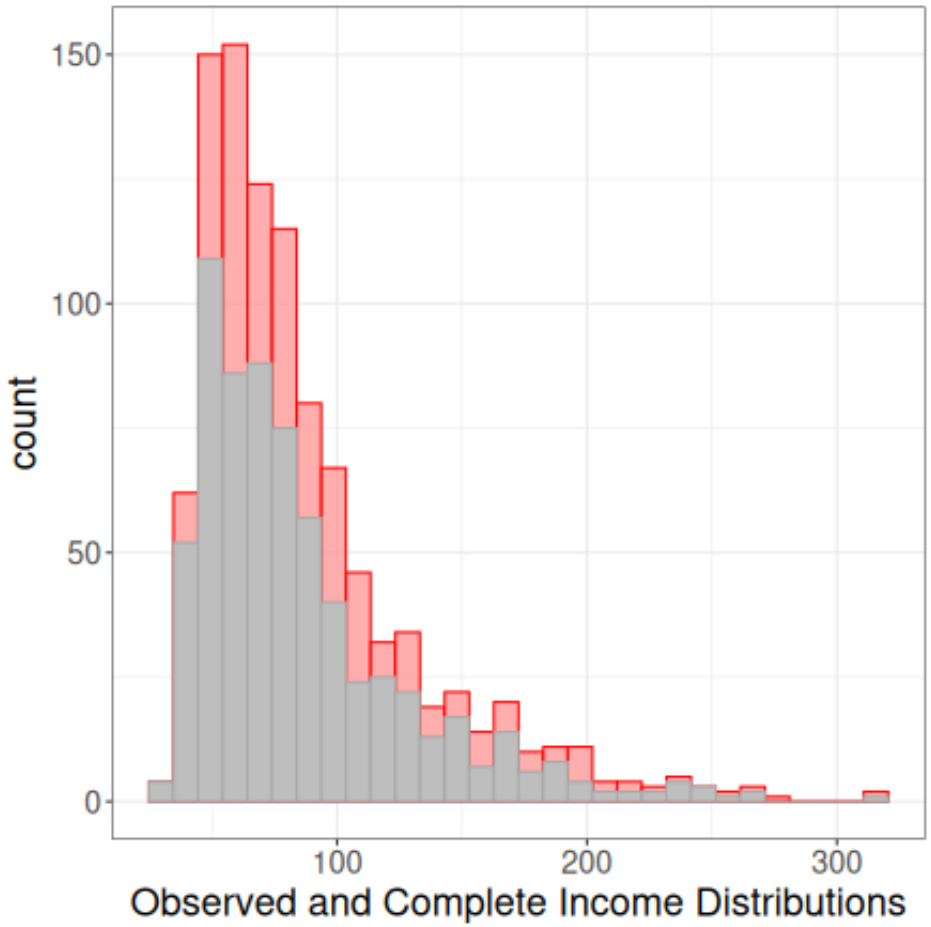
Example: Estimating Average Income

Suppose the complete income distribution for n=1000 households with a householder age 25+ is:



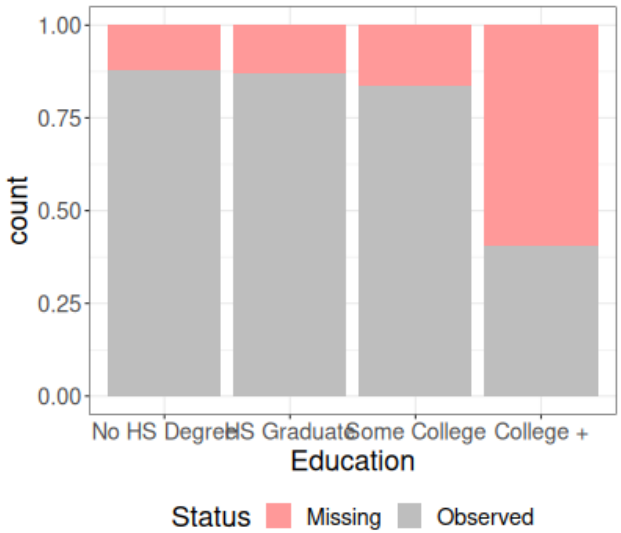
Example: Estimating Average Income

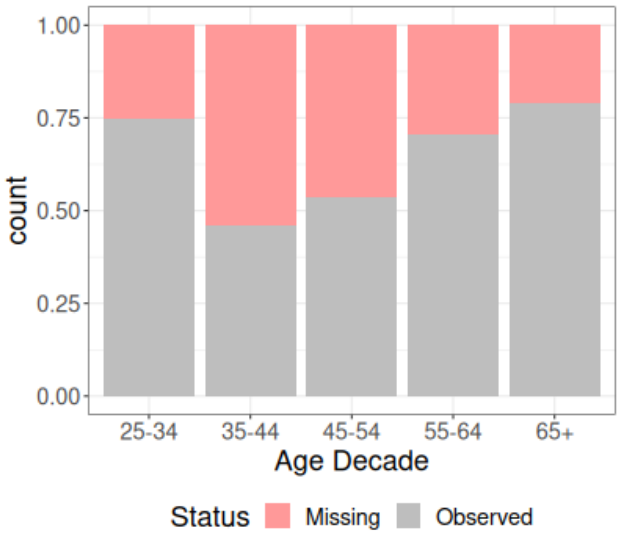
Unfortunately, only n=668 of the households replied to the survey:

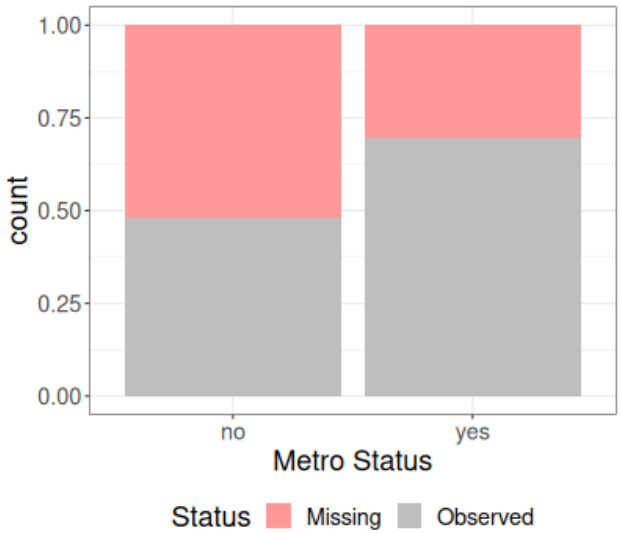


Example: Estimating Average Income

It was also a seemingly-nonrandom 668 households:







Example: Estimating Average Income

```
> mean(sim.census$incomeTot)
[1] 87.15229
> mean(sim.census$obsIncome, na.rm=T)
[1] 85.41959
```

- Full-data mean: \$87,152
- Using observed data (*listwise deletion*): \$85,420
- Question: Can we better estimate the mean if we *weight* the observations?

Example: Estimating Average Income

- Missingness model: $\text{logit}\{P(C = 1|\text{Educ}, \text{Age}, \text{Metro})\} = X\beta$

```
> ## Missingness model (C=1 if observed)
> missmod <- glm(C ~ educ + age + metro, data=census, family="binomial")
> summary(missmod)

Call:
glm(formula = C ~ educ + age + metro, family = "binomial", data = census)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.9960     0.2963  -3.361 0.000776 ***
educHS            2.6078     0.2363  11.038  < 2e-16 ***
educNoHS          2.6580     0.3558   7.471 7.95e-14 ***
educSomeCollege   2.2941     0.2232  10.279  < 2e-16 ***
age35-44         -1.6031     0.2849  -5.626 1.84e-08 ***
age45-54         -1.7002     0.3037  -5.599 2.16e-08 ***
age55-64         -0.6881     0.2833  -2.428 0.015163 *
age65+            0.1729     0.2669   0.648 0.517131
metroyes          1.3837     0.2290   6.041 1.53e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Estimating Average Income

- Weighted analysis model (just estimating a mean):

```
> ## Analysis model
> obsprobs <- predict(missmod, type="response")  ## Probability of observed
> anmod <- lm(obsIncome ~ 1, data=census, weights = 1/obsprobs)
> anmod

Call:
lm(formula = obsIncome ~ 1, data = census, weights = 1/obsprobs)

Coefficients:
(Intercept)
      91.49
```

- IPW estimate is \$91,500 – over-estimated in this case

Example: Estimating Average Income

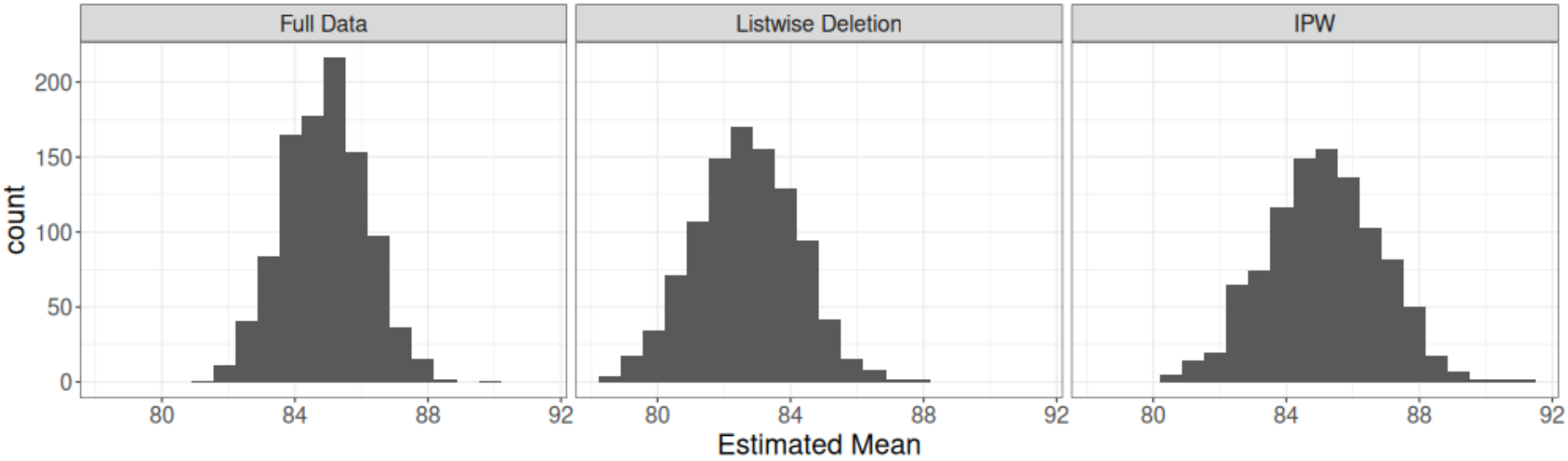
We can bootstrap (resample our incomplete census dataset) to get an estimate of the standard error:

```
> ## Bootstrapping for standard error
> ipw.ests <- c()
> for (i in 1:1000) {
+   set.seed(i)
+   census.star = sample_n(census, size=nrow(census), replace = T)
+   missmod.star <- glm(C ~ educ + age + metro, data=census.star, family="binomial")
+   obsprobs.star <- predict(missmod.star, type="response")
+   anmod.star <- lm(obsIncome ~ 1, data=census.star, weights = 1/obsprobs.star)
+
+   ipw.ests <- c(ipw.ests, coef(anmod.star)[1])
+ }
> mean(ipw.ests)
[1] 91.37983
> sd(ipw.ests)
[1] 3.969893
```

With bootstrapping, we estimate mean income is \$91,380 with SE \$3,970.

Estimating Average Income

Let's check long-term behavior with repeated simulations (in thousand $):



Method	Mean	Bootstrap SD($\bar{y}$)	95% CI Coverage
Full-data	84.9	1.27	95.2%
Listwise deletion	82.7	1.55	69.3%
IPW	85.1	1.73	95.6%

From Means to Regression

- So far, we have only estimated an unconditional mean, $\mu = E(Y)$, with auxiliary variables V
- Now suppose we have an outcome Y, covariates of interest X, and auxiliary variables V
- We are interested in a linear regression model, $E(Y|X = x) = X\beta$
  - Note: semiparametric, as we haven't specified joint distribution of X, Y
- We assume the variables in V are needed for MAR
  - $P(C = 1|Y, X, V) = P(C = 1|X, V) = \pi(X, V)$

IPW for Regression

- Proceeds almost exactly as before
- Complete-case OLS estimator solves the estimating equation

$$\sum_{i=1}^{N} C_i X_i^\top \{Y_i - X_i^\top \beta\} = 0$$

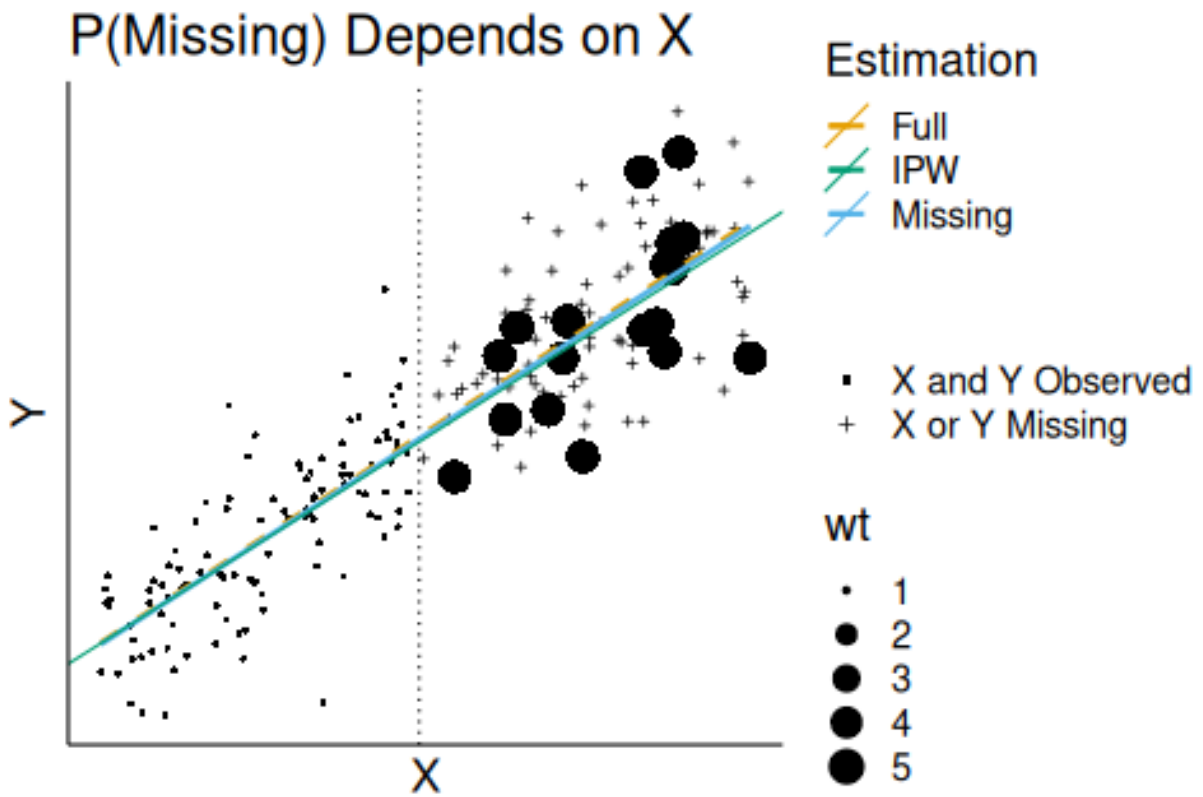
  - Not an unbiased estimating equation (proof similar to above)

- IPW estimating equation,

$$\sum_{i=1}^{N} \frac{C_i}{\pi(X_i, V_i; \hat{\phi})} X_i^\top (Y_i - X_i^\top \beta) = 0,$$

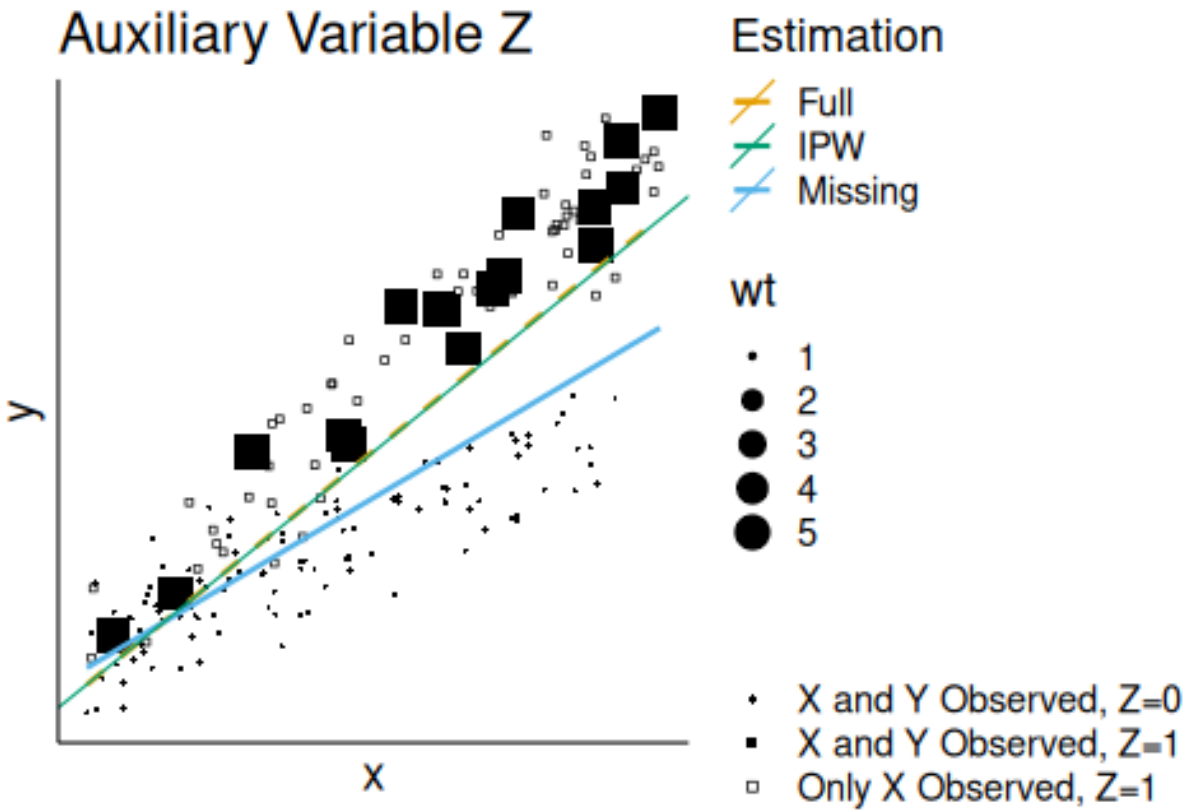
  yields consistent estimates

IPW for Regression: Simulations



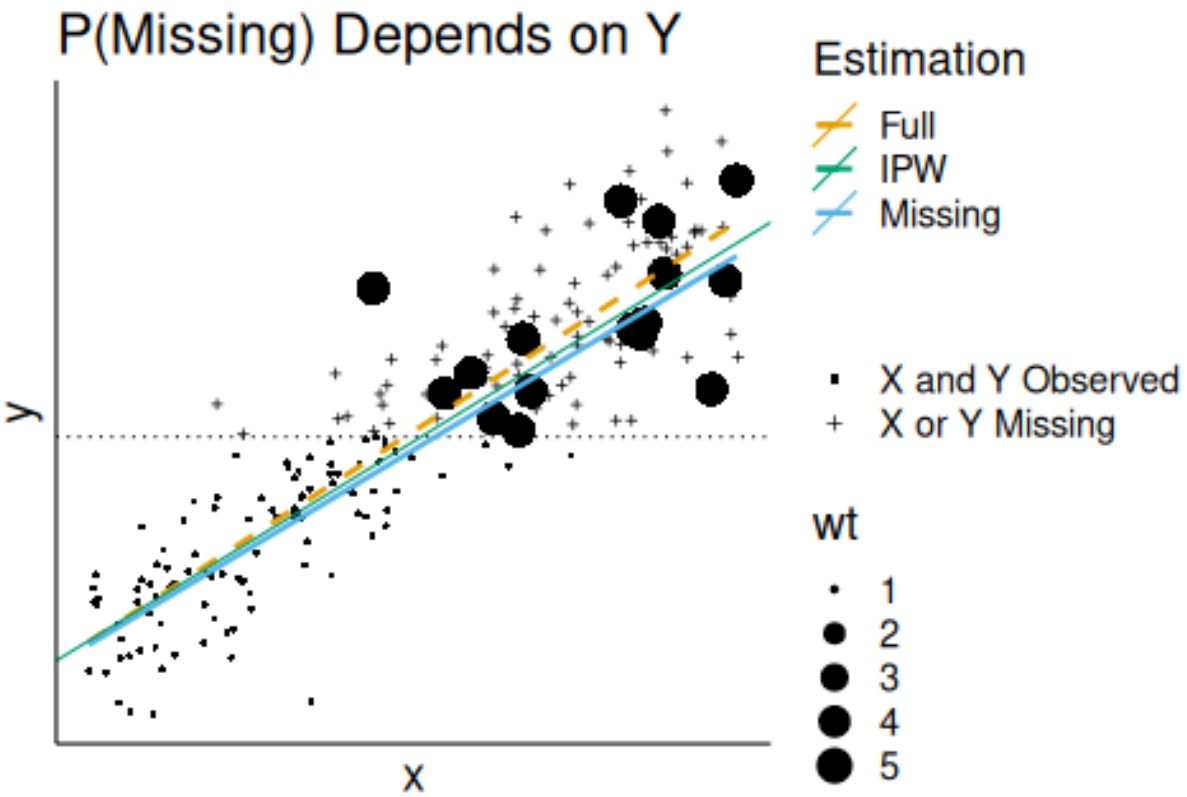
- (MCAR: every observation gets equal weight)
- MAR where missingness depends on X-variable included in model: already unbiased for (unweighted) complete-case analysis

IPW for Regression: Simulations



- MAR where missingness in depends on auxiliary Z in missingness model but not analysis model: listwise is biased, IPW unbiased

IPW for Regression: Simulations



- MAR in X depending on Y: listwise is biased, IPW unbiased

IPW for Regression: Simulations

Based on 1,000 simulated datasets of size n = 500 with 40% missingness:

Setting	Method	Bias	SD($\hat{\beta}$)	95% CI Coverage
MCAR	CC	0.003	0.058	94%
	IPW (model SE)	0.003	0.058	94%
	IPW (robust SE)			94%
Missing Y|modeled X_2	CC	0.003	0.056	96%
	IPW (model SE)	0.006	0.078	85%
	IPW (robust SE)			95%
Missing Y|auxiliary X_2	CC	-0.055	0.057	83%
	IPW (model SE)	0.003	0.076	83%
	IPW (robust SE)			94%
Missing $X\|Y$	CC	-0.081	0.056	67%
	IPW (model SE)	0.003	0.076	86%
	IPW (robust SE)			95%

What About Missing Covariates?

- Most common use case for IPW is missing outcome data (with fully observed covariates)
- IPW can also be used to weight cases with a missing key predictor X_1
  - Then the missingness model is for missingness in X_1 rather than Y
  - May depend on fully observed $X_2, ..., X_p, Y$ (but not X_1)
- With monotone missingness in covariates, can iteratively condition on each covariate to estimate $P(C)$
- With non-monotone missing covariates, options include:
  - Combine MI (for missing covariates) and IPW (to weight cases)
  - More complex models – e.g., Markov randomized monotone missing model (Robins and Gill, 1997)

What Could Go Wrong?

- The results above rely on correct specification of the missingness model, which requires:
  - Class of model is correct, e.g., logistic regression
  - Necessary predictors are included and treated correctly (e.g., transformations and interactions)
- Very large weights are often a red flag
  - Larger weights = smaller effective sample size
  - More often due to model misspecification than to genuinely low P(complete case)
- Approaches to handling very large weights:
  - Truncation
    - Try a few cut-off values as a sensitivity analysis
  - Try semi-parametric models – e.g., kernel smoothers
    - Requires few predictors
  - Try other models (e.g., robit regression, penalized logistic regression)
    - Less readily available in software

Case Study: NHANES

- Desired model:

$$\text{TotChol} = \beta_0 + \beta_1 \times \text{Age} + \beta_2 \times \text{Pulse} + \beta_3 \times \text{BPSysAve}$$

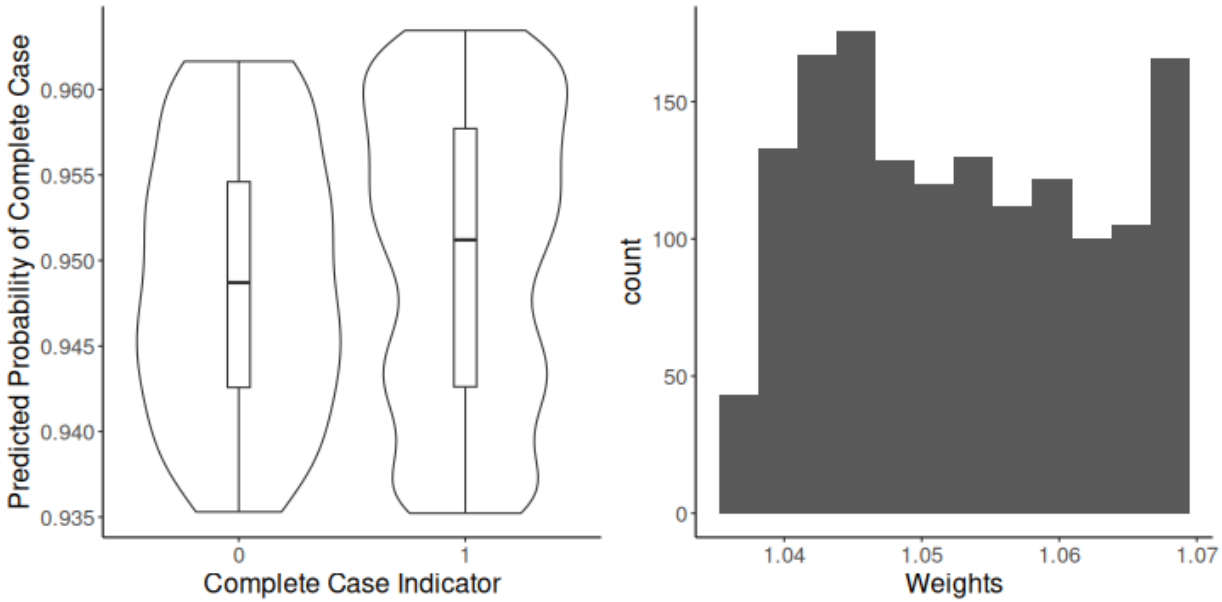
- Monotone missingness: BPSysAve has about 5% missing data (Pulse is missing for all but 2 of these subjects)

Case Study: NHANES

Step 1a: fit missingness model (possibly using auxiliary variables)

```
> # Complete-case indicator
> nha$C <- as.numeric(!is.na(nha$BPSysAve10))
>
> # Logistic regression for P(complete observation)
> miss_mod <- glm(C ~ Age10 + TotChol, data = nha, family = binomial(link="logit"))
>
> # Predicted probabilities and weights
> nha$pred_probs <- predict(miss_mod, type="response")
> nha$weights <- 1/nha$pred_probs
```

Step 1b: check weight distribution



Case Study: NHANES

Step 2a: fit analysis model using weights

```
> # Fit model
> ipw_mod <- lm(TotChol ~ Age10 + PulseSec + BPSysAve10,
+               weights = weights, data = nha)
> model.res <- coef(summary(ipw_mod))
```

Step 2b: Robust or bootstrap standard errors

```
> # Obtain robust standard errors (lmtest package with SE functions from sandwich)
> robust.res <- coeftest(ipw_mod, vcov = vcovHC(ipw_mod, type = "HC3"))
>
> # Obtain bootstrap standard errors
> boot_ipw_ests <- matrix(nrow=1000, ncol=4)
> for (i in 1:1000) {
+     set.seed(i)
+     nha.star <- sample_n(nha, size=nrow(nha), replace = T)
+     missmod.star <- glm(C ~ Age10 + TotChol, data = nha.star,
+                         family = binomial(link="logit"))
+     obsprobs.star <- predict(missmod.star, type="response")
+     ipwmod.star <- lm(TotChol ~ Age10 + PulseSec + BPSysAve10,
+                       weights = 1/obsprobs.star, data = nha.star)
+     boot_ipw_ests[i,] <- coef(ipwmod.star)
+ }
> boot.se <- apply(boot_ipw_ests, 2, sd)
```

Case Study: NHANES

Results:

- Bootstrap SEs are (slightly) smaller than robust SEs
  - Robust SE tends to be conservative
- Both are noticeably larger than model-based SEs
  - Model-based SEs after IPW are generally not valid (too small)

Coefficient	Estimate	Model SE	Robust SE	Bootstrap SE
Intercept	2.729	0.265	0.304	0.304
Age (per 10 years)	0.301	0.032	0.032	0.032
Pulse (beats/6 sec)	0.035	0.022	0.024	0.023
Systolic BP (per 10 mmHg)	0.075	0.019	0.024	0.023

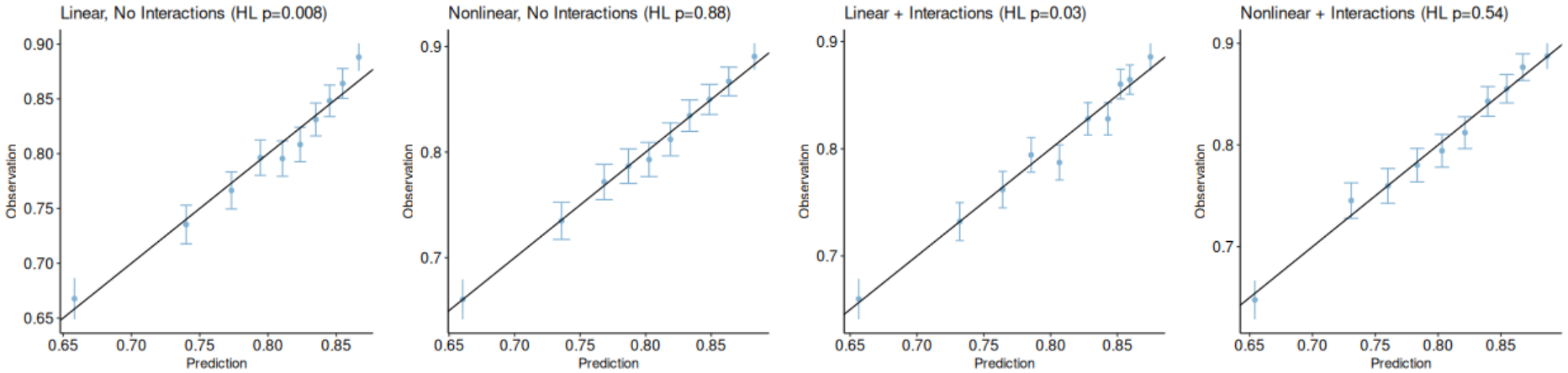
Case Study: NHANES

$$E[\text{Total Cholesterol}|\text{Covariates}] = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Pulse} + \beta_3 \text{Syst BP}$$

	LD	**ML**	**MI**	**IPW**
Intercept	2.73 (0.26)	2.68 (0.26)	2.73 (0.27)	2.73 (0.30)
Age (per 10 yrs)	0.30 (0.03)	0.31 (0.03)	0.32 (0.03)	0.30 (0.03)
Pulse (per 6 sec)	0.035 (0.022)	0.035 (0.022)	0.033 (0.022)	0.035 (0.023)
Syst BP (per 10 mmHg)	0.076 (0.020)	0.076 (0.020)	0.073 (0.020)	0.075 (0.023)

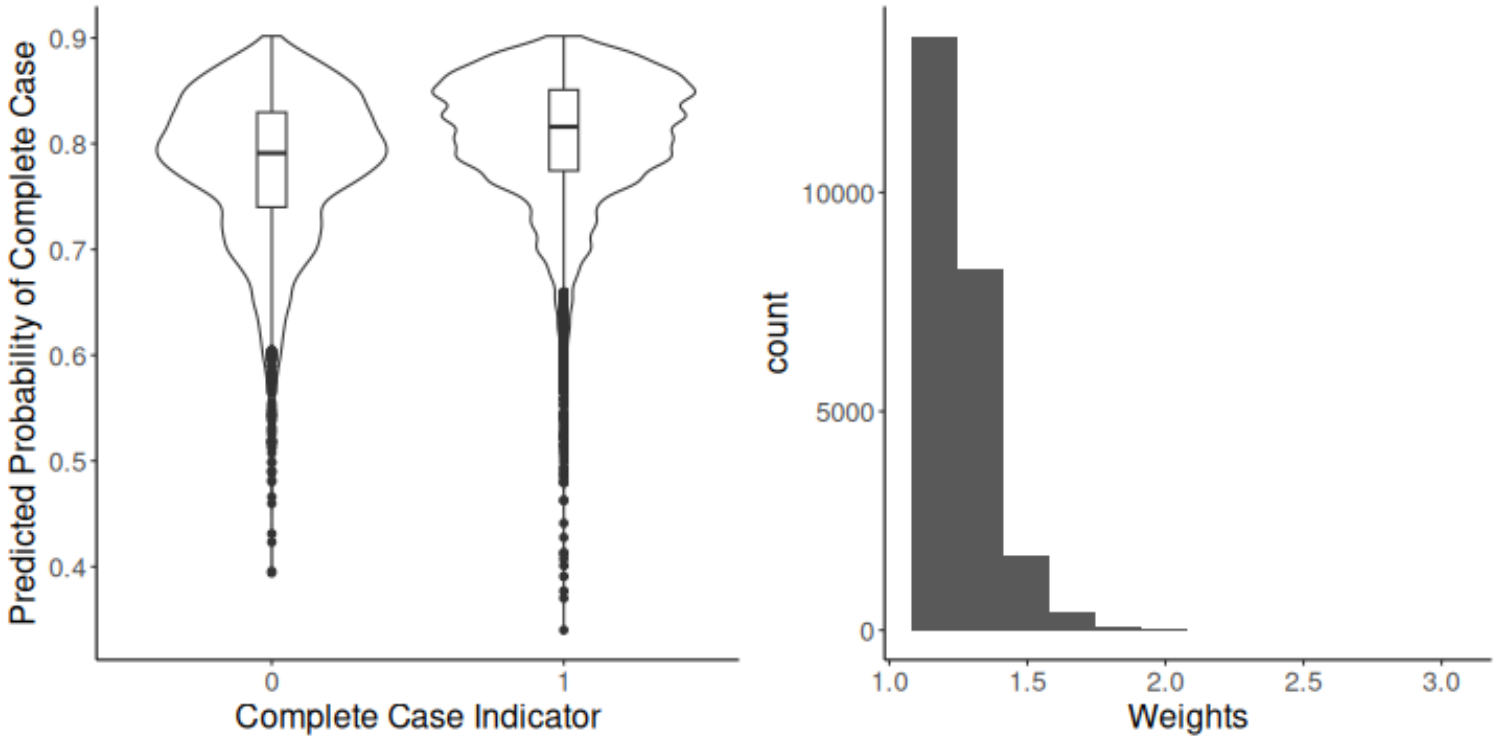
Case Study: NACC

- Goal: investigate whether patient's APOE4 status modifies the relationship between risk factor and MMSE score
- Step 1: Decide on a missingness model
  - Logistic regression with outcome = APOE observation status
  - Predictors: risk factors and MMSE, with/without (1) all pairwise interactions, (2) flexible age and MMSE associations (GAM)
    - Seems like allowing nonlinear age and MMSE associations helps
    - (Calibration plots and Hosmer-Lemeshow test as evidence – don't treat as infallible)



Case Study: NACC

- Check weight distribution for extreme weights (further indicator of missingness model misspecification)
- Looks okay in this case



Case Study: NACC

IPW Results (robust SE)

	No copies	**At least one**	**P-value**
Age (per 10 yrs)	-0.38 (0.06)	-1.09 (0.09)	<0.0001
Sex (Female)	0.13 (0.09)	-0.02 (0.12)	0.45
Parkinson's	-1.21 (0.37)	-3.01 (0.43)	0.0005
Depression	-0.70 (0.14)	0.26 (0.15)	<0.0001
Stroke	-1.65 (0.27)	-1.89 (0.26)	0.48
Clinical AD	-5.28 (0.12)	-6.15 (0.12)	<0.0001

Case Study: NACC

Among those with no copies of APOE4 allele:

	LD	ML	MI	IPW
Age (per 10 yrs)	-0.36 (0.06)	-0.37 (0.06)	-0.38 (0.06)	-0.38 (0.06)
Sex (Female)	0.19 (0.09)	0.22 (0.09)	0.22 (0.09)	0.13 (0.09)
Parkinson's	-0.94 (0.31)	-0.89 (0.28)	-0.89 (0.28)	-1.21 (0.37)
Depression	-0.62 (0.13)	-0.70 (0.12)	-0.71 (0.13)	-0.70 (0.14)
Stroke	-1.46 (0.20)	-1.51 (0.20)	-1.57 (0.20)	-1.65 (0.27)
Clinical AD	-5.07 (0.10)	-5.02 (0.10)	-5.03 (0.11)	-5.28 (0.12)

Case Study: NACC

Among those with at least one copy of APOE4 allele:

	LD	ML	MI	IPW
Age (per 10 yrs)	-1.02 (0.08)	-1.20 (0.08)	-1.20 (0.08)	-1.09 (0.09)
Sex (Female)	0.08 (0.11)	-0.02 (0.11)	-0.02 (0.11)	-0.02 (0.12)
Parkinson's	-2.58 (0.46)	-2.87 (0.45)	-2.87 (0.44)	-3.01 (0.43)
Depression	0.25 (0.15)	0.25 (0.15)	0.25 (0.15)	0.26 (0.15)
Stroke	-1.72 (0.26)	-1.90 (0.26)	-1.80 (0.26)	-1.89 (0.26)
Clinical AD	-5.94 (0.11)	-6.41 (0.11)	-6.42 (0.11)	-6.15 (0.12)

Case Study: NACC

Significance of Effect Modification (p-values):

	LD	**ML**	**MI**	**IPW**
Age (per 10 yrs)	<0.0001	<0.0001	<0.0001	<0.0001
Sex (Female)	0.45	0.11	0.11	0.45
Parkinson's	0.003	0.0005	0.0003	0.0005
Depression	<0.0001	<0.0001	<0.0001	<0.0001
Stroke	0.43	0.28	0.51	0.48
Clinical AD	<0.0001	<0.0001	<0.0001	<0.0001

Improving on IPW Complete Case Analysis

- IPW is still a complete case method (valid, but inefficient)
- We can do better by *augmenting* the estimating equation with information about partially observed cases

Augmented Inverse Probability Weighting (aIPW)

- The original IPW estimating equation for $\mu = E(Y)$ was

$$\sum_{i=1}^{N} \frac{C_i}{\pi(V_i; \hat{\phi})}(Y_i - \mu) = 0$$

- Consider augmenting with the expected residuals:

$$\sum_{i=1}^{N} \left[\frac{C_i}{\pi(V_i; \hat{\phi})}(Y_i - \mu) + \left(1 - \frac{C_i}{\pi(V_i; \hat{\phi})}\right) E\{(Y_i - \mu)|V_i\} \right] = 0$$

  - For observed Y $(C_i = 1)$: $\frac{1}{\pi(V_i;\hat{\phi})}(Y_i - \mu) + \frac{\pi(V_i;\hat{\phi})-1}{\pi(V_i;\hat{\phi})} E\{(Y_i - \mu)|V_i\}$, i.e., weighted combination of observed and expected residual
  - For missing Y $(C_i = 0)$: $E\{(Y_i - \mu)|V_i\}$

Augmented Inverse Probability Weighting (aIPW)

- This leads to the estimator

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{C_i Y_i}{\pi(V_i;\hat{\phi})} - \frac{C_i - \pi(V_i;\hat{\phi})}{\pi(V_i;\hat{\phi})}E(Y_i|V_i)\right\}$$

- Assuming MAR, we can estimate $E(Y|V)$ using observed data only, i.e., $E(Y|V) = E(Y|V, C = 1)$

- If we fit a regression model to estimate $E(Y|V, C = 1) = m(v;\xi)$, then the final aIPW estimator is

$$\hat{\mu}^{aipw} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{C_i Y_i}{\pi(V_i;\hat{\phi})} - \frac{C_i - \pi(V_i;\hat{\phi})}{\pi(V_i;\hat{\phi})}m(V_i;\hat{\xi})\right\}$$

Properties of aIPW Estimator

- aIPW estimator is doubly robust: consistent if **either**
  - The missingness model $\pi(v;\phi)$ is correctly specified, or
  - The model $m(v;\xi)$ for $E(Y|V)$ is correctly specified
- Also more efficient than IPW estimator
- (Practical challenge: most implementations apply in limited settings – e.g., for particular causal inference estimands, allowing missingness only in outcome, requiring key predictor to be binary)

Example: Estimating Average Income

- R package `iWeigReg` can perform DR estimation for a mean
- Missingness model: $\text{logit}\{P(C = 1|\text{Educ}, \text{Age}, \text{Metro})\} = X\beta$
- Outcome regression model: $E(\text{Income}|\text{Educ}, \text{Age}, \text{Metro}) = X\beta$

```
library(iWeigReg)
y = sim.census$obsIncome
y[is.na(y)] <- 0
eta1.glm <- glm(obsIncome ~ metro + age + educ, data=sim.census, subset=C==1)
eta1.hat <- predict.glm(eta1.glm, newdata=sim.census, type="response")

ppi.glm <- glm(C ~ metro + age + educ, data=sim.census, family=binomial)
X <- model.matrix(ppi.glm)
ppi.hat <- ppi.glm$fitted.values
```

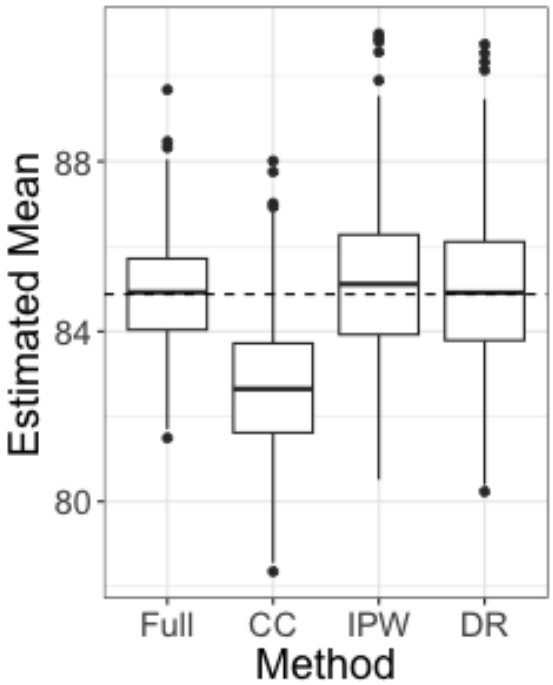
Example: Estimating Average Income

- Complete-data mean: \$87,150
- Listwise deletion mean: \$85,420
- IPW estimate: \$91,370 (SE = \$3,970)
- DR estimate is \$88,860 (SE = \$3,880)

```
> res <- mn.creg(y = y,
+          tr = census$C,
+          g = eta1.hat,
+          p = ppi.hat,
+          X = X,
+          evar = TRUE)
> res$mu
[1] 88.86302
> res$v
[1] 15.01584
```

Example: Estimating Average Income

Let's check long-term behavior with repeated simulations (in thousand $):



Method	Mean	SD($\bar{y}$)	95% CI Coverage
Full-data	84.9	1.23	95%
Listwise deletion (CC)	82.7	1.54	69%
IPW	85.1	1.72	96%
DR	84.9	1.67	95%

aIPW for Regression

- Augmenting the IPW estimating equations for regression follows a similar procedure:
  - Define analysis model, $\mu(x;\theta) = E(Y|X;\theta)$
  - Define augmented model, $m(x,v;\xi) = E(Y|X,V;\xi)$
  - Math proceeds as before

- Note: $m(x,v;\xi)$ and $\mu(x;\theta)$ must be **compatible**, i.e.,

$$\mu(X;\theta) = E(Y|X) = E\{E(Y|X,V)|X\} = E\{m(X,V;\xi)|X\}$$

  - One solution is to assume centered residual and centered V are multivariate normal (conditional on X)

Weighting in Longitudinal Studies

Weighted generalized estimating equations (WGEE) can account for dropout under MAR. Briefly:

- Model the probability of remaining in the study at time j given presence at time $j-1$ (e.g., using logistic regression),

$$\pi_{ij} = P(C_{ij} = 1 | C_{i,j-1} = 1)$$

- Then these conditional probabilities can be chained to get the unconditional P(observed at time j),

$$P(\text{complete until j}) = \pi_{ij} \times \pi_{i,j-1} \times \cdots \times \pi_{i1}$$

- So, the weight becomes:

$$w_{ij} = [\pi_{ij} \times \pi_{i,j-1} \times \cdots \times \pi_{i1}]^{-1}$$

Weighting in Longitudinal Studies

- These weights are combined with weights related to the "working covariance" in the GEE framework for estimation
  - Beyond the scope of this module – see, e.g., Applied Longitudinal Analysis by Fitzmaurice, Laird, and Ware
  - R package `wgeesel` includes `wgee` for regular IPW (weighted GEE) and `drgee` for doubly-robust GEE estimation

- Assumptions: typical GEE assumptions, plus model for P(dropout) must be correct

Summary: IPW

- **Characteristics**
  - Semi-parametric: solution to unbiased estimating equations
  - Weights complete cases by the probability of being a complete case
  - Missingness model is based on fully observed characteristics
- **Advantages**
  - Consistent, assuming (1) MAR, (2) correctly specified mean model, and (3) correctly specified missingness model
  - May be easier to specify missingness model than full joint likelihood
- **Disadvantages**
  - Inefficient: only uses complete cases
  - Variance estimation can be conservative (especially if there are very large weights) – consider alternatives such as bootstrapping
- **Practical notes**
  - Model misspecification may cause very large weights; check for this

Summary: aIPW

- **Characteristics**
  - Adds a second term to the estimating equation (weighted combination of observed and expected residual)
  - Allows some information from incomplete cases to be used
- **Advantages**
  - More efficient than IPW (uses more information)
  - **Doubly robust**: consistent if *either* missingness model *or* model for $E(Y|V)$ is correctly specified
- **Disadvantages**
  - May be difficult to specify the two key pieces; unclear, for example, whether to aim for relatively minimalist or maximal models
  - May still be less efficient than maximum likelihood methods
  - Convenient, reliable, general-case software implementations are hard to find

MI vs. IPW

- MI is typically more efficient and allows more missing data patterns
- There are a few special cases where IPW might still be preferred:
  - Specification of the imputation model can be difficult (e.g., including all appropriate nonlinear terms and interactions)
  - If most cases with missing data have many missing variables, it may make sense to exclude those cases rather than make assumptions about the joint distribution of all missing variables
- Combining MI and IPW:
  - Define a rule for including an individual (e.g., based on % data observed)
  - Impute missing values for included individuals
  - Use IPW to account for exclusion of individuals who did not meet criteria

(See Seaman et al (2012), Little, Carpenter, and Lee (2022), and others for more)

Overview

The Difficult Case: Non-Ignorable Missingness

"All happy families are alike; each unhappy family is unhappy in its own way." –Leo Tolstoy, *Anna Karenina*

All of the analyses we have seen so far require researchers to make assumptions that cannot be verified based on the observed data.

- Are the data MAR? Would need to know $P(R|Z^{full}) = P(R|Z^{obs})$
- What is the MNAR mechanism? Would need to know $P(R|Z^{full})$

So, we will assess robustness to departures from our assumptions (i.e., sensitivity of conclusions to different forms of dependence on Z^{full}).

Problem of Identifiability

- The ideal full data would be (R, Z)
- The observed data, (R, Z^{obs}), are a **many-to-one** transformation
  - There are many joint distributions $p_{R,Z}$ that will lead to the same distribution for (R, RZ)
  - We cannot determine which of these actually generated the data
- Idea of sensitivity analysis:
  - Specify distribution of $R|Z$ (or $Z|R$) $\implies$ full joint distribution becomes identifiable
  - See how much estimates and inference change

Approach 1: Pattern-Mixture Models

- Sensitivity parameter (δ) specifies how distribution of Z differs between missing and non-missing observations
  - Natural connection to MI and likelihood-based methods

- Joint (R, Z) likelihood is factored as follows:

$$f(z_i, r_i|\gamma) = f(z_i|r_i, \gamma) \times f(r_i)$$
$$= \text{model for } z \text{ within pattern } r_i \times \text{probability of pattern } r_i$$

Approach 1: Pattern-Mixture Models

- E.g., in the case of simple mean estimation, assume

$$E(Z|R=1) = E(Z|R=0) + \delta$$

and plug in a plausible range of δ

- In the case of more complex models, may use mean shift,

$$E(Z|V,R=1) = E(Z|V,R=0) + \delta$$

or a functional that depends on V,

$$E(Z|V,R=1) = E(Z|V,R=0) + s(V;\delta)$$

Example: NHANES Average Systolic Blood Pressure

Mean estimation:

- Suppose we are interested in estimating average systolic blood pressure (5% missingness)
- Observed average: $\bar{x}^{CC} = 110.9$
- We'll consider a range of mean shifts for subjects with missing data:

$$E(Z|R=1) = E(Z|R=0) + \delta$$

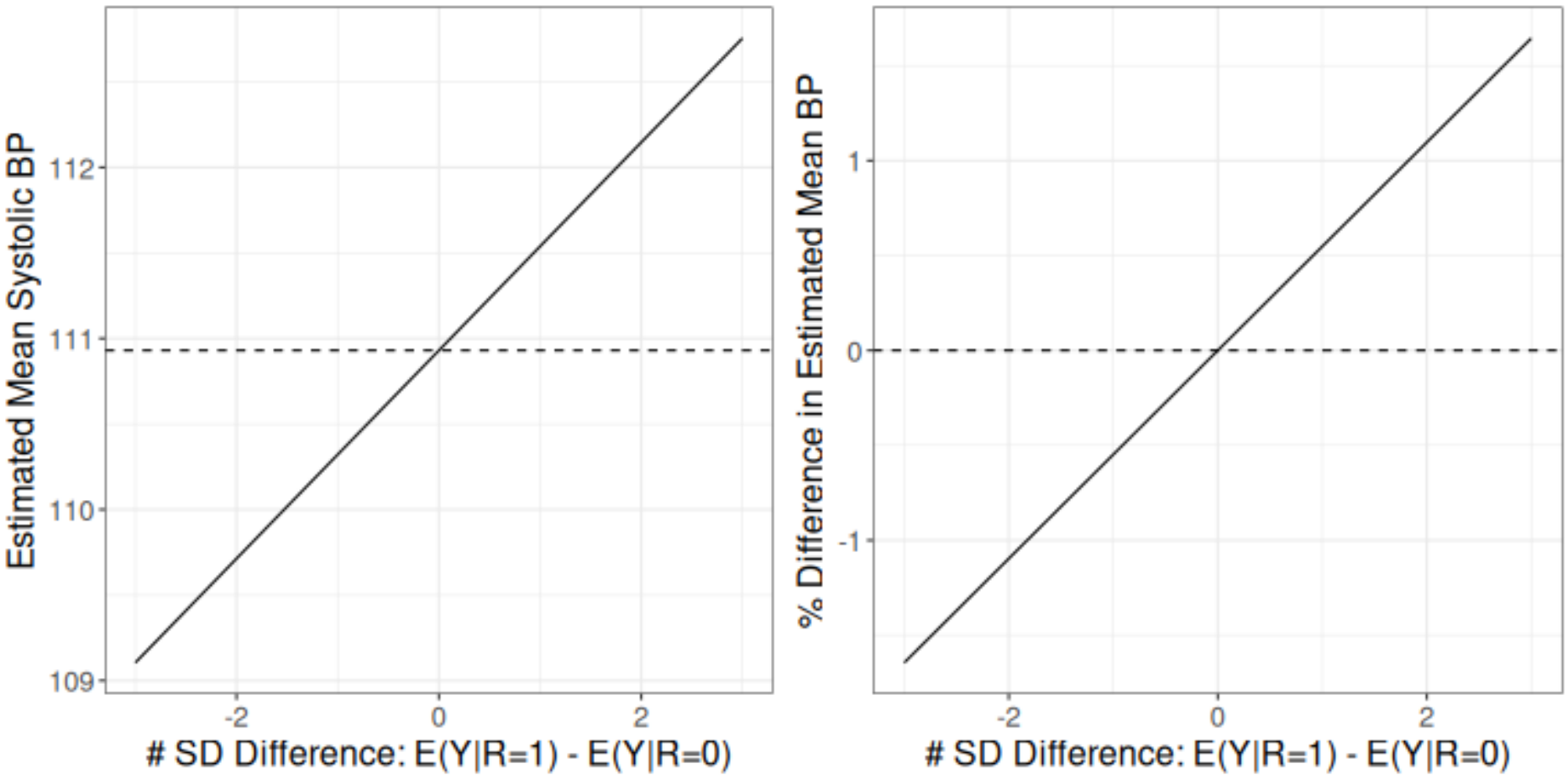
$$\delta \in \{-3, -2, -1, +1, +2, +3SD\}$$

Example: NHANES Average Systolic Blood Pressure

```
> # Observed mean
> cc.xbar <- mean(nha$BPSysAve, na.rm=T) ; cc.xbar
[1] 110.93
>
> # Find percent missing and SD
> pct.miss <- mean(is.na(nha$BPSysAve)); pct.miss
[1] 0.0499002
> s <- sd(nha$BPSysAve, na.rm=T) ; s
[1] 12.19122
>
> # Pattern mixture model for mean: +/- 1, 2, 3 SD
> xbars <- (1-pct.miss) * cc.xbar +
+     pct.miss * (cc.xbar + seq(from=-3, to=3, by=1) * s)
> xbars
[1] 109.1049 109.7133 110.3216 110.9300 111.5383 112.1467 112.7550
```

Example: NHANES Average Systolic Blood Pressure

- We can visualize this as either absolute or percent change in the estimated mean depending on severity of MNAR:



Example: NHANES Regression for Cholesterol

Mean estimation:

- Now let's return to the example from earlier: we want to fit the model

$$\text{TotChol} = \beta_0 + \beta_1 \times \text{Age} + \beta_2 \times \text{Pulse} + \beta_3 \times \text{BPSysAve}$$

- The `mice` package used for multiple imputation can also perform sensitivity analyses using pattern mixture models (argument `post`)
- We'll consider the same MNAR patterns: observations missing SBP have a linear shift of $\delta \in \{-3, -2, -1, +1, +2, +3SD\}$ from the value we would have predicted

Example: NHANES Regression for Cholesterol

For Step 1, we need to specify the function to apply to imputed missing values (for each variable for which we want to do a sensitivity analysis):

```
> # Pattern mixture model in imputation: +/- 1, 2, 3 SD
> delta.all <- seq(from=-3, to=3, by=1) * s  ; delta.all
[1] -36.57365 -24.38244 -12.19122   0.00000  12.19122  24.38244  36.57365
>
> # Multiple imputation for all deltas
> imp.all <- vector("list", length(delta.all))  # set up output list for imputations
> post <- ini$post  # initializes post-adjustment vector
> for (i in 1:length(delta.all)) {
+   this.d <- delta.all[i]
+   # command for adjusting estimates post-imputation by adding delta
+   cmd <- paste("imp[[j]][,i] <- imp[[j]][,i] +", this.d)
+   # specify that cmd should be applied to sysBP variable
+   post["BPSysAve"] <- cmd
+   # Run imputation with post-adjustment
+   imp.all[[i]] <- mice(nha2, post=post, m=10, maxit=5, seed=i, print=FALSE)
+ }
```

Example: NHANES Regression for Cholesterol

Steps 2 and 3 look the same as before:

```
> # Step 2: fit linear models for all 10 imputed datasets (in each imputation setting)
> step2.mods <- lapply(imp.all, FUN = function(imp) {
+   with(imp, lm(TotChol ~ Age + Pulse + BPSysAve))
+ })
>
> # Step 3: pool results
> step3.res <- lapply(step2.mods, FUN = function(mods) pool(mods))
```

Example: NHANES Regression for Cholesterol

The pooled results for β_3, the coefficient for systolic BP, are:

Delta	Estimate	d	SE	P-value
-3 SD	0.0053	-0.0024	0.0016	0.0011
-2 SD	0.0064	-0.0013	0.0018	0.0002
-1 SD	0.0074	-0.0003	0.0019	0.0001
0 SD (MAR)	0.0077	0	0.0019	0.0001
1 SD	0.0074	-0.0003	0.0019	0.0002
2 SD	0.0067	-0.0010	0.0018	0.0002
3 SD	0.0052	-0.0025	0.0016	0.0013

where $d = \hat{\beta}(\delta) - \hat{\beta}(\delta = 0)$

Approach 2: Selection Models

- Sensitivity parameter (δ) specifies how probability of $R = 1$ differs depending on value of Z
- Joint (R, Z) likelihood is factored as follows:

$$f(z_i, r_i|\theta, \phi) = f(z_i|\theta) \times f(r_i|z_i, \phi)$$
$$= \text{complete-data model} \times \text{model for md mechanism}$$

Approach 2: Selection Models

- E.g., assume $\text{logit}\{P(C = 1|Z)\} = \alpha + \delta Z$
  - e^{δ} is the odds ratio for not missing associated with a unit change in Z:

$$e^{\delta} = \frac{\text{Odds}(C = 1|Z = z + 1)}{\text{Odds}(C = 1|Z = z)}$$

  - Larger (positive or negative) values for δ indicate larger departures from MCAR
  - (Economists often use probit instead for this)

- Goal: find the range of $\mu(\delta)$ induced from the distribution of the observed data for a plausible range of δ

Approach 2: Selection Models

- We can't estimate α and δ from the observed data (Z is missing whenever $C = 0$)

- Instead, we'll fix δ and estimate α
  - Let $\pi(Z; \alpha, \delta) = P(C = 1|Z) = \text{expit}(\alpha + \delta Z)$
  - This may be rewritten as

$$\frac{1}{\pi(Z; \alpha, \delta)} = 1 + \exp\{-(\alpha + \delta Z)\}$$

$$\implies 1 = \pi(\delta)\left[1 + \exp\{-(\alpha + \delta Z)\}\right]$$

so $\hat{\alpha}(\delta)$ may be estimated by solving the estimating equation

$$\sum_{i=1}^{N} \left(C_i[1 + \exp\{-(\alpha + \delta Z_i)\}] - 1\right) = 0$$

Approach 2: Selection Models

- For estimating the mean μ, with the assumed δ and estimated α, is:

$$\hat{\mu}(\delta) = \frac{1}{N}\sum_{i=1}^{N} C_i Z_i \left[1 + \exp\{-(\hat{\alpha}(\delta) + \delta Z_i)\}\right]$$

(note the second term is effectively an IPW weight: $\frac{1}{\pi(Z;\alpha,\delta)}$)

- In the case of more complex models, assume

$$\text{logit}\{P(C = 1|Z, V)\} = \alpha(V) + \delta(Z, V)$$

for sensitivity function $\delta(Z, V)$

  - From there, analagous to the procedure above

Pattern-Mixture Models vs. Selection Models

- Pattern-mixture models:
  - Conceptually natural if population strata are defined by missing-data pattern
  - May be easier to implement – closer to current form of data
  - Avoids specifying form of missing-data mechanism
  - Conceptually more similar to imputation (fill in missing data, assumptions about MNAR)

- Selection models:
  - Conceptually natural if interested in entire population
  - (Used to be) more common in the literature
  - Sensitive to the selected form of missing-data mechanism
  - May pair naturally with IPW (weight observed data, assumptions about MNAR)
  - Out-of-the-box software (in R) is much harder to find and less well-documented for selection models than for pattern-mixture models

Overview

Case Studies: Missing Data in the Literature

- Missing data has existed for as long as data has existed, and people have been doing *something* about it
- We have discussed some effective methods, and some less effective methods, for handling missing data
- What are people choosing to do in practice?
  - Case studies from recent literature in epidemiological and medical journals
  - For each, think about:
    1. What assumptions are they making with that choice?
    2. What would you want to know in order to evaluate whether those assumptions are reasonable?
    3. Do you expect this approach to yield valid inference?

Ex: Occupational Sitting and Standing (AJE)

American Journal of Epidemiology

Vol. 187, No. 1
DOI: 10.1093/aje/kwx298
Advance Access publication:
August 11, 2017

Original Contribution

The Relationship Between Occupational Standing and Sitting and Incident Heart Disease Over a 12-Year Period in Ontario, Canada

Peter Smith*, Huiting Ma, Richard H. Glazier, Mahée Gilbert-Ouimet, and Cameron Mustard

* Correspondence to Dr. Peter Smith, Institute for Work and Health, 481 University Avenue, Suite 800, Toronto M5G 2E9, Ontario, Canada (e-mail psmith@iwh.on.ca).

Initially submitted March 30, 2017; accepted for publication August 3, 2017.

While a growing body of research is examining the impacts of prolonged occupational sitting on cardiovascular and other health risk factors, relatively little work has examined the effects of occupational standing. The objectives of this paper were to examine the relationship between occupations that require predominantly sitting and those that require predominantly standing and incident heart disease. A prospective cohort study combining responses to a population health survey with administrative health-care records, linked at the individual level, was conducted in Ontario, Canada. The sample included 7,320 employed labor-market participants (50% male) working 15 hours a week or more

Ex: Occupational Sitting and Standing (AJE)

Outline of study:

- Objective: to compare the association of occupational sitting vs. occupational standing with incident heart disease
- Data: administrative medical records for 7,320 respondents to the Canadian Community Health Survey (50% male)
  - Complete cases only (86% of eligible sample)
- Conclusions: occupational standing is associated with 50% higher risk of incident heart disease than occupational sitting

Ex: Occupational Sitting and Standing (AJE)

Of the original sample of 8,873 respondents, 350 (4%) either reported having preexisting heart disease or were captured in the Ontario Myocardial Infarction Database or the Ontario Congestive Heart Failure Database prior to the interview date, and were removed, leaving a sample of 8,523 respondents. Of this sample, 562 respondents were missing information on work exposures, with an additional 641 respondents missing information on sociodemographic characteristics, health measures, or health behaviors, leaving a final analytical sample of 7,320 respondents (50% male), which is 86% of the eligible sample. A logistic regression analysis examined variables associated with the probability of missing work exposures, and missing sociodemographic, health, or health-behavior measures. Men were more likely than women to be missing work-exposure information, while women, respondents in urban locations, and those working in other body positions were more likely to be missing sociodemographic, health, or health-behavior measures. No relationship was found between age and having missing information on work exposures or having missing information on sociodemographic, health, or health-behavior measures.

- Complete case analysis
- Checked MAR vs MCAR (association of missing indicator with covariates)

Ex: Cardiovascular Risk Score (BMJ)

BMJ

RESEARCH

Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study

Julia Hippisley-Cox, professor of clinical epidemiology and general practice,[1] Carol Coupland, senior lecturer in medical statistics,[1] Yana Vinogradova, research fellow in medical statistics,[1] John Robson, senior lecturer in general practice,[2] Margaret May, research fellow in medical statistics,[3] Peter Brindle research and development strategy lead[4]

[1]Tower Building, University Park, Nottingham NG2 7RD

[2]Centre for Health Sciences, Queen Mary's School of Medicine and Dentistry, London

[3]Department of Social Medicine, University of Bristol

ABSTRACT

Objective To derive a new cardiovascular disease risk score (QRISK) for the United Kingdom and to validate its performance against the established Framingham cardiovascular disease algorithm and a newly developed Scottish score (ASSIGN).

according to the Framingham algorithm. UK estimates for 2005 based on QRISK give 3.2 million patients aged 35-74 at high risk, with the Framingham algorithm predicting 4.7 million and ASSIGN 5.1 million. Overall, 53 668 patients in the validation dataset (9% of the total) would be reclassified from high to low risk or vice versa using

Ex: Cardiovascular Risk Score (BMJ)

Outline of study:

- Objective: to derive and validate a new UK cardiovascular disease risk score (QRISK)
- Data: 1.28 million patients (318 practices) to develop score, 0.61 million patients (160 practices) to validate
- Conclusions: New score does not overestimate cardiovascular disease risk nearly as much as Framingham or a Scottish score

Ex: Cardiovascular Risk Score (BMJ)

tested for interactions between systolic blood pressure and antihypertensive treatment and between smoking and deprivation.

Initially our models were fitted using patients without any missing data (complete case analysis), and the fractional polynomial terms were obtained from the complete case analysis using a multivariable approach. However, because patients with complete data might have a different health status and risk of cardiovascular disease compared with those with missing data, we fitted our principal models on the basis of multiple imputed datasets using Rubin's rules to combine effect estimates and estimate standard errors.[24 25] We used the ICE procedure in Stata[26] to obtain five imputed datasets.

Ex: Cardiovascular Risk Score (BMJ)

The authors revised their imputation procedure shortly after publication (in response to letters from readers):

Rapid response to:

Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study

BMJ 2007 ; 335 doi: https://doi.org/10.1136/bmj.39261.471806.55 (Published 19 July 2007)
Cite this as: *BMJ* 2007;335:136

- "We have improved our implementation of multiple imputation for missing data. Our revised model for imputation includes more variables in addition to those originally included. [...] the estimated hazard ratios for the cholesterol ratio term in males and females increased to more biologically plausible values given our combined outcome of cardiovascular disease."

Ex: Cardiovascular Risk Score (BMJ)

Their original analysis was later included in an overview of potential pitfalls of MI (and revised analysis highlighted as an appropriate correction):

BMJ. 2009; 338: b2393.
Published online 2009 Jun 29. doi: 10.1136/bmj.b2393
PMCID: PMC2714692
PMID: 19564179

Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls

Jonathan A C Sterne, professor of medical statistics and epidemiology,[1] Ian R White, senior scientist,[2] John B Carlin, director of clinical epidemiology and biostatistics unit,[3] Michael Spratt, research associate,[1] Patrick Royston, senior scientist ,[4] Michael G Kenward, professor of biostatistics,[5] Angela M Wood, lecturer in biostatistics,[6] and James R Carpenter, reader in medical and social statistics[5]

Most studies have some missing data. **Jonathan Sterne and colleagues** describe the appropriate use and reporting of the multiple imputation approach to dealing with them

Ex: Cardiovascular Risk Score (BMJ)

- "The researchers correctly identified a difficulty with missing data in their database and used multiple imputation to handle the missing data in their analysis. In their published prediction model, however, cardiovascular risk was found to be unrelated to cholesterol (coded as the ratio of total to high density lipoprotein cholesterol), which was surprising."

- "The main reasons for the unexpected finding of a null association between cholesterol level and cardiovascular risk were omission of the cardiovascular disease outcome when imputing missing cholesterol values and calculation of the ratio of cholesterol to HDL based on imputed cholesterol and HDL values, which led to extreme values of the ratio being included in estimations. The impact of these pitfalls was increased by the high proportion of missing data (70% of HDL cholesterol values were missing)."

Ex: Telehealth in Psychiatric Care (JAMA Psychiatry)

Original Investigation FREE

August 25, 2021

Comparison of Teleintegrated Care and Telereferral Care for Treating Complex Psychiatric Disorders in Primary Care

A Pragmatic Randomized Comparative Effectiveness Trial

John C. Fortney, PhD[1,2]; Amy M. Bauer, MS, MD[1]; Joseph M. Cerimele, MPH, MD[1]; et al

» Author Affiliations | Article Information

JAMA Psychiatry. 2021;78(11):1189-1199. doi:10.1001/jamapsychiatry.2021.2318

Ex: Telehealth in Psychiatric Care (JAMA Psychiatry)

Outline of study:

- Objective: to compare "two clinic-to-clinic interactive video approaches" to delivering care for PTSD or bipolar disorder
  - Treatment 1 (TCC): Telepsychiatry = remote delivery of direct psychiatric care
  - Treatment 2 (TER): Collaborative care = psychiatrist consults remotely with primary care team
- Participants and approach:
  - 1004 patients from 24 primary care clinics
  - Randomized to TCC or TER, followed for 12 months
  - Primary outcome: patient-reported outcome, Veterans RAND 12-item Health Survey Mental Component Summary score
- Conclusions: significant improvement in both groups, no significant difference between the groups

Ex: Telehealth in Psychiatric Care (JAMA Psychiatry)

- 71-72% of each group were still in the study at 6 month follow-up
- 63-64% were still in the study at 12 months

the delta method.

To account for missing data at follow-up, 100 complete data sets were imputed[31] using random forest imputation.[32-34] Imputation was stratified by initial randomization group to allow potentially different effects and separate covariance structures by group.[35] Potential missingness mechanisms were examined by correlating loss to follow-up status with key baseline characteristics. To gauge the sensitivity of results to violations of the missing-at-random assumption, we calculated how the primary outcome estimate changed under different proportions and effect sizes of nonignorable missingness. Statistical analyses were conducted using Mplus version 8 (Muthen & Muthen).

Ex: Telehealth in Psychiatric Care (JAMA Psychiatry)

eTable. Dropout analysis

	Observed at 12 months	Missing at 12 months	Cohen's *d* or % difference	p-value
n	638 (63.6%)	366 (36.5%)		
Condition = Telepsychiatry Enhanced Referral (%)	313 (49.1%)	183 (50.0%)	-0.9%	.77
Gender (%)				.72
Female	454 (71.2%)	249 (68.0%)	3.2%	---
Male	175 (27.4%)	110 (30.1%)	-2.7%	---
Transgender, non-binary, or other	7 (1.1%)	5 (1.4)	0.3%	---
Age	40.8 (13.2)	36.9 (11.8)	-0.31	<.001
Minority status (%)	213 (33.4%)	127 (34.7%)	-1.3%	.68
Income above poverty threshold (%)	223 (37.4%)	103 (29.5%)	7.9%	.014
Positive Bipolar Disorder screen (%)	219 (34.3%)	148 (40.4%)	-6.1%	.05
Baseline MCS (mean (SD))	22.1 (10.0)		0.02	.79
Baseline PTSD Symptoms - PCL-5 (mean (SD))	47.2 (18.2)	48.8 (17.0)	-0.09	.18
Baseline Depression - SCL-20 (mean (SD))	2.4 (0.7)	2.5 (0.7)	-0.03	.68
Baseline Altman Mania Rating Scale (mean (SD))	9.3 (3.3)	9.7 (3.6)	-0.14	.046

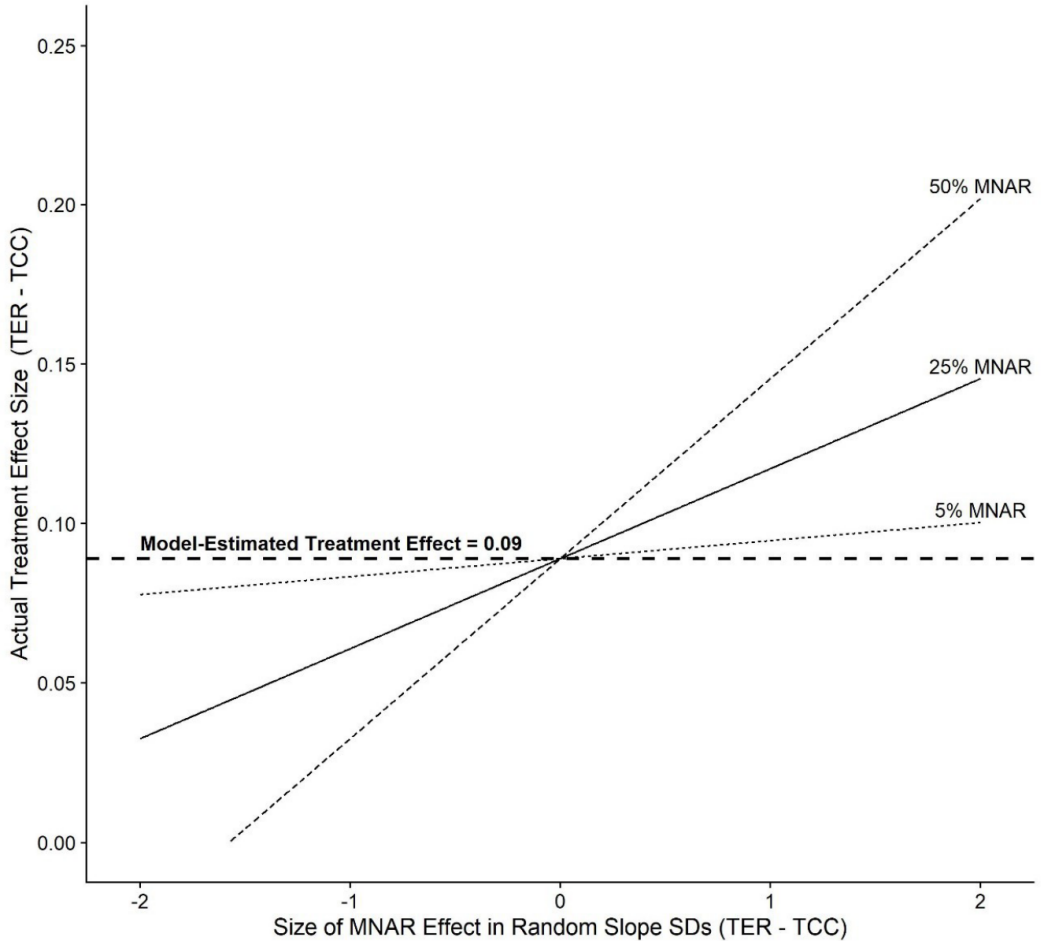
Note. P-values were not adjusted for multiple comparisons.

(Supplementary Table)

Ex: Telehealth in Psychiatric Care (JAMA Psychiatry)

"Even under the most extreme assumptions, where 50% of missing data were MNAR and the observed treatment effect was biased downward because participants missing in TER had trajectories 2 standard deviations better than those missing in TCC, bias would remain very small (d = 0.11)."

eFigure 2. Sensitivity of treatment effects across size and proportion of MAR violations



(Supplementary Figure 2)

ELSEVIER

Check for updates

Journal of Clinical Epidemiology 134 (2021) 79–88

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework

Katherine J. Lee[a,b,*], Kate M. Tilling[c], Rosie P. Cornish[c], Roderick J.A. Little[d], Melanie L. Bell[e], Els Goetghebeur[f], Joseph W. Hogan[g], James R. Carpenter[h], on behalf of the STRATOS initiative

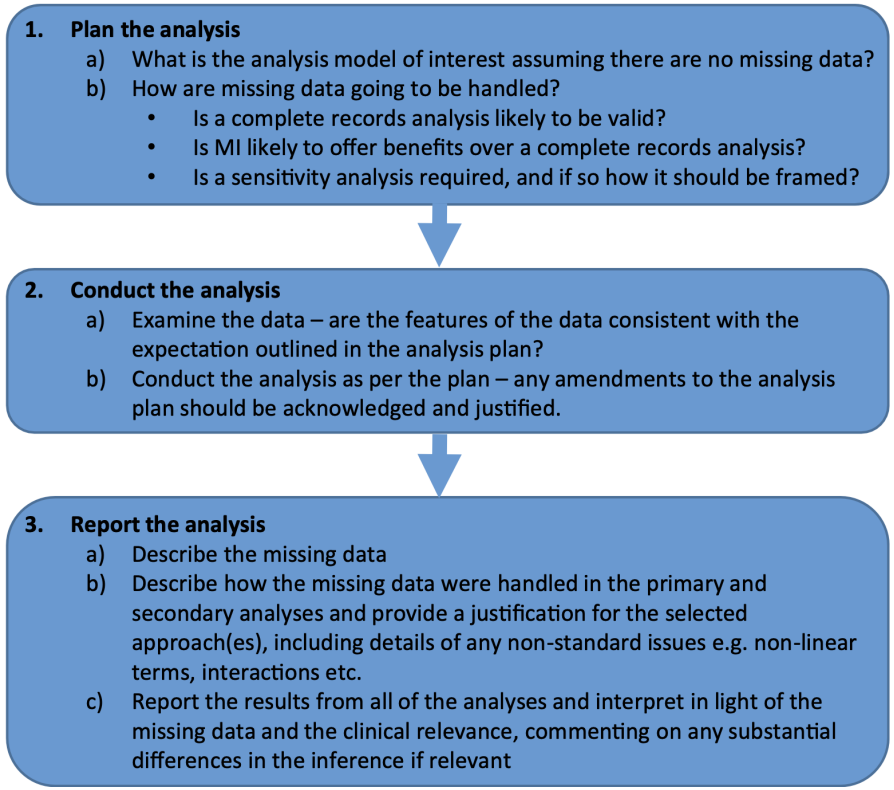


Fig. 1. The framework.

Reporting

STROBE

Strengthening the reporting of observational studies in epidemiology

Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding
		(*b*) Describe any methods used to examine subgroups and interactions
		(*c*) Explain how missing data were addressed
		(*d*) *Cohort study*—If applicable, explain how loss to follow-up was addressed *Case-control study*—If applicable, explain how matching of cases and controls was addressed *Cross-sectional study*—If applicable, describe analytical methods taking account of sampling strategy
Results		
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed
		(b) Give reasons for non-participation at each stage
		(c) Consider use of a flow diagram
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders
		(b) Indicate number of participants with missing data for each variable of interest
		(c) *Cohort study*—Summarise follow-up time (eg, average and total amount)

Template - Multiple Imputation

Template suggestion from Stef van Buuren:

The percentage of missing values across the nine variables varied between 0 and 34%. In total 1601 out of 3801 records (42%) were incomplete. Many girls had no score because the nurse felt that the measurement was "unnecessary," or because the girl did not give permission. Older girls had many more missing data. We used multiple imputation to create and analyze 40 multiply imputed datasets. Methodologists currently regard multiple imputation as a state-of-the-art technique because it improves accuracy and statistical power relative to other missing data techniques. Incomplete variables were imputed under fully conditional specification, using the default settings of the `mice` 3.0 package (Van Buuren and Groothuis-Oudshoorn 2011). The parameters of substantive interest were estimated in each imputed dataset separately, and combined using Rubin's rules. For comparison, we also performed the analysis on the subset of complete cases.

Advice I

- No free lunch! We have to make assumptions about the missingness mechanism (unless missingness was by design)
  - It is worthwhile to spend effort in the study design stages to limit missingess and plan to collect variables that can be used to predict missing values
  - An assumption of MCAR is usually not appropriate
  - Use background knowledge, scientific literature, etc. to defend assumptions
- Carefully plan ahead of time and describe in detail the procedure(s) taken to handling missing data in analysis plans and manuscripts

Advice II

- Complete case analysis can sometimes be valid and optimal(!)
  - Example: regression of Y on $X \rightarrow$ complete case is valid if missingness (in Y or X) depends on X only and not Y (not always efficient, however)
- Approaches that exist for handling missing data assuming MAR (that we discussed here) include maximum likelihood, multiple imputation, and inverse probability weighting
  - Efficiency
  - Ease of implementation
- Use sensitivity analyses to assess robustness of inference

References and Resources - Textbooks

- Allison, P. D. (2002). Missing data. https://doi.org/10.4135/9781412985079
- Enders, C. K. (2022). Applied missing data analysis.
  - https://www.appliedmissingdata.com/ contains code and other resources
- Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data. https://https://doi.org/10.1002/9781119013563
- Van Buuren, S. (2018). Flexible imputation of missing data.
  - Available online: https://stefvanbuuren.name/fimd/

References and Resources - Articles

- Carpenter, J. R., & Smuk, M. (2021). Missing data: A statistical framework for practice. *Biometrical Journal*. https://doi.org/10.1002/bimj.202000196
- Lee, K. J., Tilling, K.M., ... & Carpenter, J. R. (2021). Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. *Journal of Clinical Epidemiology*. https://doi.org/10.1016/j.jclinepi.2021.01.008
- Little, R. J., Carpenter, J. R., & Lee, K. J. (2022). A comparison of three popular methods for handling missing data: complete-case analysis, inverse probability weighting, and multiple imputation. *Sociological Methods & Research*. https://doi.org/10.1177/00491241221113873
- Little, R. J., D'Agostino, ... & Stern, H. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*. https://doi.org/10.1056/NEJMsr1203730

References and Resources - Articles

- Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*. https://doi.org/10.1037/1082-989X.7.2.147
  - Enders, C. K. (2023). Missing data: An update on the state of the art. *Psychological Methods*. https://doi.org/10.1037/met0000563
- Seaman, S. R., & White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*. https://doi.org/10.1177/096228021039574

Overview

Appendix: EM Algorithm Example

Preliminaries:

$$f(y_i|x_{1i}, x_{2i}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}))^2\right)$$

$$f(x_{2i}|x_{1i}) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2}(x_{2i} - (\alpha_0 + \alpha_1 x_{1i}))^2\right)$$

$$\begin{aligned} l(\theta|x_{1i}, x_{2i}, y_i) &= -1/2 \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}))^2 \\ &\quad - 1/2 \log(2\pi\tau^2) - \frac{1}{2\tau^2}(x_{2i} - (\alpha_0 + \alpha_1 x_{1i}))^2 \\ &= -1/2 \log(2\pi\sigma^2) - 1/2 \log(2\pi\tau^2) \\ &\quad - \frac{1}{2\sigma^2}\left[(y_i - (\beta_0 + \beta_1 x_{1i}))^2 - 2\beta_2(y_i - (\beta_0 + \beta_1 x_{1i}))x_{2i} + \beta_2^2 x_{2i}^2\right] \\ &\quad - \frac{1}{2\tau^2}\left[x_{2i}^2 - 2(\alpha_0 + \alpha_1 x_{1i})x_{2i} + (\alpha_0 + \alpha_1 x_{1i})^2\right] \end{aligned}$$

Appendix: EM Algorithm Example

Thus, $l(\theta|\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}) = \sum_{i=1}^{n} l(\theta|x_{1i}, x_{2i}, y_i)$

$$E[l(\theta|\mathbf{x}_1, \mathbf{x}_2, \mathbf{y})|\mathbf{x}_2^{(obs)}, \theta = \theta^{(t-1)}] = \sum_{i=1}^{n} E[l(\theta|x_{1i}, x_{2i}, y_i)|x_2^{(obs)}, \theta = \theta^{(t-1)}]$$

$$= -n/2 \log(2\pi\sigma^2) - n/2 \log(2\pi\tau^2)$$

$$- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i}))^2 - \frac{1}{2\tau^2} \sum_{i=1}^{n} (\alpha_0 + \alpha_1 x_{1i})^2$$

$$+ \frac{1}{\sigma^2} \sum_{i=1}^{n} \beta_2 (y_i - (\beta_0 + \beta_1 x_{1i})) E[X_{2i}|\mathbf{x}_2^{(obs)}, \theta = \theta^{(t-1)}]$$

$$- \frac{1}{2\sigma^2} \sum_{i=1}^{n} \beta_2^2 E[X_{2i}^2|\mathbf{x}_2^{(obs)}, \theta = \theta^{(t-1)}]$$

$$- \frac{1}{2\tau^2} E[X_{2i}^2|\mathbf{x}_2^{(obs)}, \theta = \theta^{(t-1)}]$$

$$+ \frac{1}{\tau^2} \sum_{i=1}^{n} (\alpha_0 + \alpha_1 x_{1i}) E[X_{2i}|\mathbf{x}_2^{(obs)}, \theta = \theta^{(t-1)}]$$

Appendix: EM Algorithm Example

Note that (conditional on x_1)

$$\begin{pmatrix} Y_i \\ X_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_0 + \beta_1 x_{1i} + \beta_2(\alpha_0 + \alpha_1 x_{1i}) \\ \alpha_0 + \alpha_1 x_{1i} \end{pmatrix}, \begin{pmatrix} \sigma^2 + \beta_2^2\tau^2 & \beta_2\tau^2 \\ \beta_2\tau^2 & \tau^2 \end{pmatrix} \right)$$

Thus, for $i = 1, \ldots, m$ (where x_{2i} is missing)

$$E[X_{2i}|\mathbf{x}_2^{(obs)}, \theta = \theta^{(t-1)}] = \alpha_0^{(t)} + \alpha_1^{(t)} x_{1i} + \frac{\beta_2^{(t)}\tau^{2(t)}}{\beta_2^{2(t)}\tau^{2(t)} + \sigma^{2(t)}}(y_i - (\beta_0^{(t)} + \beta_1^{(t)} x_{1i} + \beta_2^{(t)}(\alpha_0^{(t)} + \alpha_1^{(t)} x_{1i}))$$

$$E[X_{2i}^2|\mathbf{x}_2^{(obs)}, \theta = \theta^{(t-1)}] = \frac{\sigma^{2(t)}}{\beta_2^{2(t)}\tau^{2(t)} + \sigma^{2(t)}} + \left(E[X_{2i}|\mathbf{x}_2^{(obs)}, \theta = \theta^{(t-1)}]\right)^2$$

otherwise, $E[X_{2i}|\mathbf{x}_2^{(obs)}, \theta = \theta^{(t-1)}] = x_{2i}$ and $E[X_{2i}^2|\mathbf{x}_2^{(obs)}, \theta = \theta^{(t-1)}] = x_{2i}^2$

Appendix: Inconsistency of CC Estimator

We can show that $\hat{\mu}^{CC}$ does not generally converge in probability to $E(Y) = \mu$:

$$\hat{\mu}^{CC} = \frac{\sum_{i=1}^{N} C_i Y_i}{\sum_{i=1}^{N} C_i} \xrightarrow{p} \frac{E(CY)}{E(C)} \qquad \text{by LLN}$$

$$= \frac{E\{E(CY|Y,V)\}}{E\{E(C|Y,V)\}} \qquad \text{by iterated expectations}$$

$$= \frac{E\{YE(C|Y,V)\}}{E\{E(C|Y,V)\}}$$

$$= \frac{E\{Y\pi(V)\}}{E\{\pi(V)\}}$$

$$\neq E(Y) = \mu \qquad \text{if } Y, V \text{ not indep.}$$

Appendix: Consistency of IPW Estimator

To prove consistency, we will show that the estimating function is unbiased, i.e.,

$$E_\mu \left\{ \frac{C}{\pi(V)}(Y - \mu) \right\} = 0$$

The proof uses the law of iterated conditional expectations:

$$\begin{aligned}
E_\mu \left\{ \frac{C}{\pi(V)}(Y - \mu) \right\} &= E_\mu \left[E \left\{ \frac{C}{\pi(V)}(Y - \mu) | Y, V \right\} \right] \\
&= E_\mu \left\{ \frac{E(C|Y, V)}{\pi(V)}(Y - \mu) \right\} \\
&= E_\mu \left\{ \frac{\pi(V)}{\pi(V)}(Y - \mu) \right\} && \text{by definition of } \pi(V) \\
&= E_\mu (Y - \mu) = 0 && \text{because } \pi(V) > 0 \text{ a.s.}
\end{aligned}$$