Dealing with Missing Data

Andrew Miles

University of Toronto Social Science
Research Methods week, 2018

Mechanisms of Missingness

- MCAR – missing completely at random
  - no pattern to the missingness
- MAR – missing at random
  - missingness depends on variables you have in your model
- MNAR – missing not at random
  - missingness tied to values of the outcome
  - (indirect) missingness tied to variables not in the model

Example

Years of Education	Political ideology	Poli Ideology (MCAR)
9	9	9
9	4	4
9	6	---
11	5	---
11	8	8
12	2	2
12	5	5
14	6	6
15	7	---

Example

Years of Education	Political ideology	Poli Ideology (MCAR)	Poli Ideology (MAR)
9	9	9	9
9	4	4	---
9	6	---	---
11	5	---	5
11	8	8	---
12	2	2	2
12	5	5	5
14	6	6	6
15	7	---	7

Example

Years of Education	Political ideology	Poli Ideology (MCAR)	Poli Ideology (MAR)	Poli Ideology (MNAR)
9	9	9	9	9
9	**4**	4	---	---
9	6	---	---	6
11	**5**	---	5	---
11	8	8	---	8
12	2	2	2	2
12	**5**	5	5	---
14	6	6	6	6
15	7	---	7	7

Why missingness is a problem

	MCAR	MAR	MNAR
larger SEs	X	X	X
biased estimates		X	X

$$\bar{y} = 5.78$$

$$\bar{y}_{MCAR} = 5.83$$

$$\bar{y}_{MAR} = 5.67$$

$$\bar{y}_{MNAR} = 6.33$$

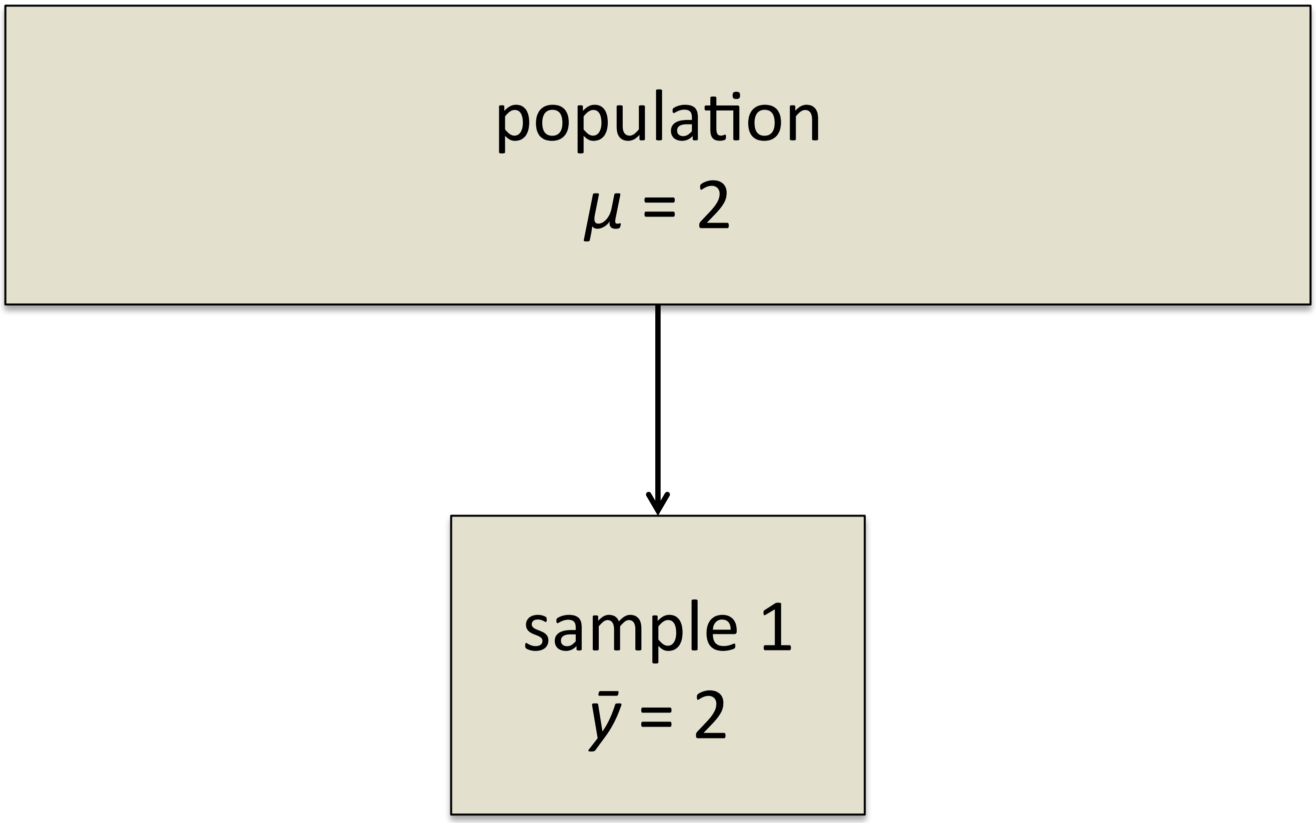
Methods for Handling Missing Data

- In the past…
  - most often: listwise deletion (still the default)
  - mean imputation, regression imputation
- Widespread consensus that two techniques are currently state-of-the-art
  - Maximum Likelihood (or Full Information ML)
  - Multiple Imputation (MI)

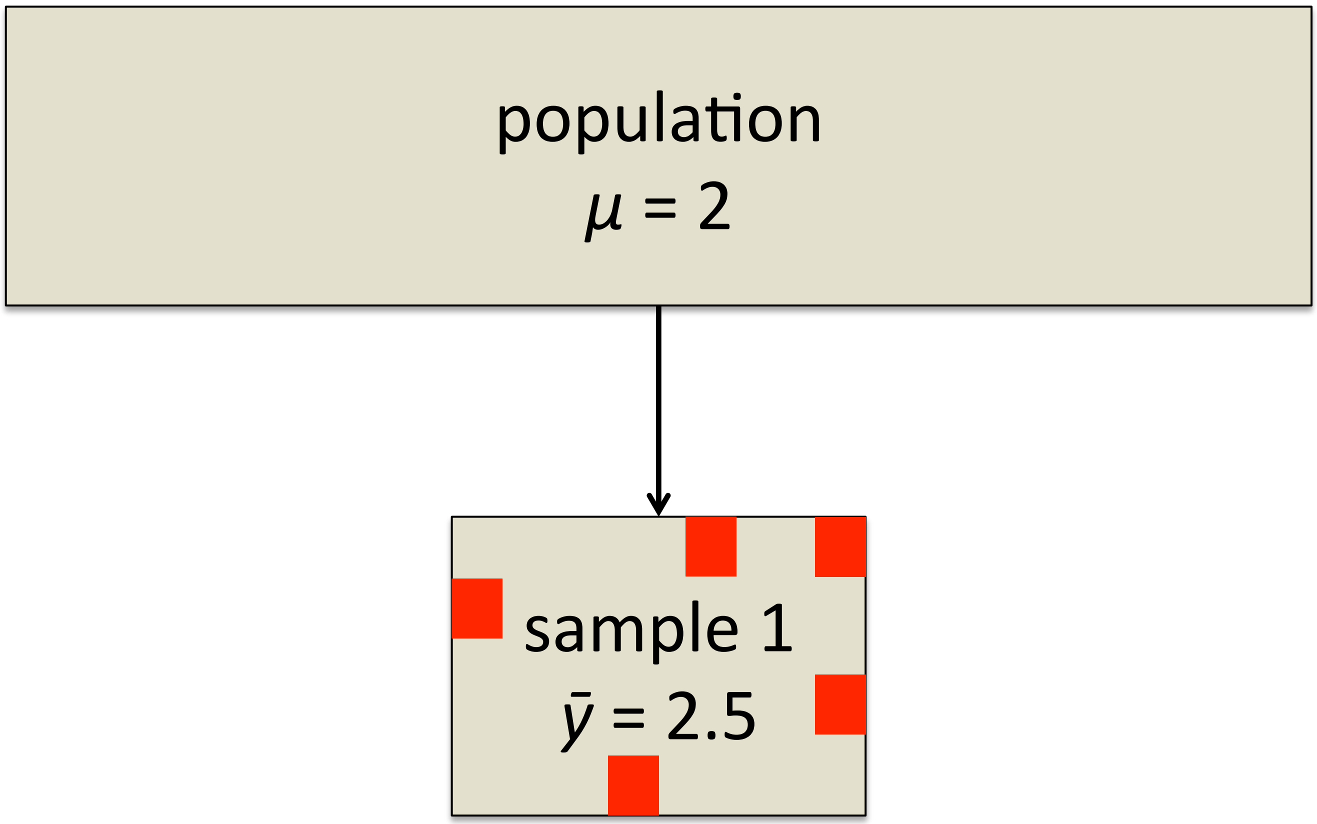
Comparing the two methods

- Both rely on MAR assumption*
- asymptotically equivalent
- FIML
  - generally simpler to use, consistent results across runs, BUT
  - only available for continuous (outcome) data
  - only implemented in structural equation modeling software
- focus will be on multiple imputation

Basic Ideas of MI



Basic Ideas of MI

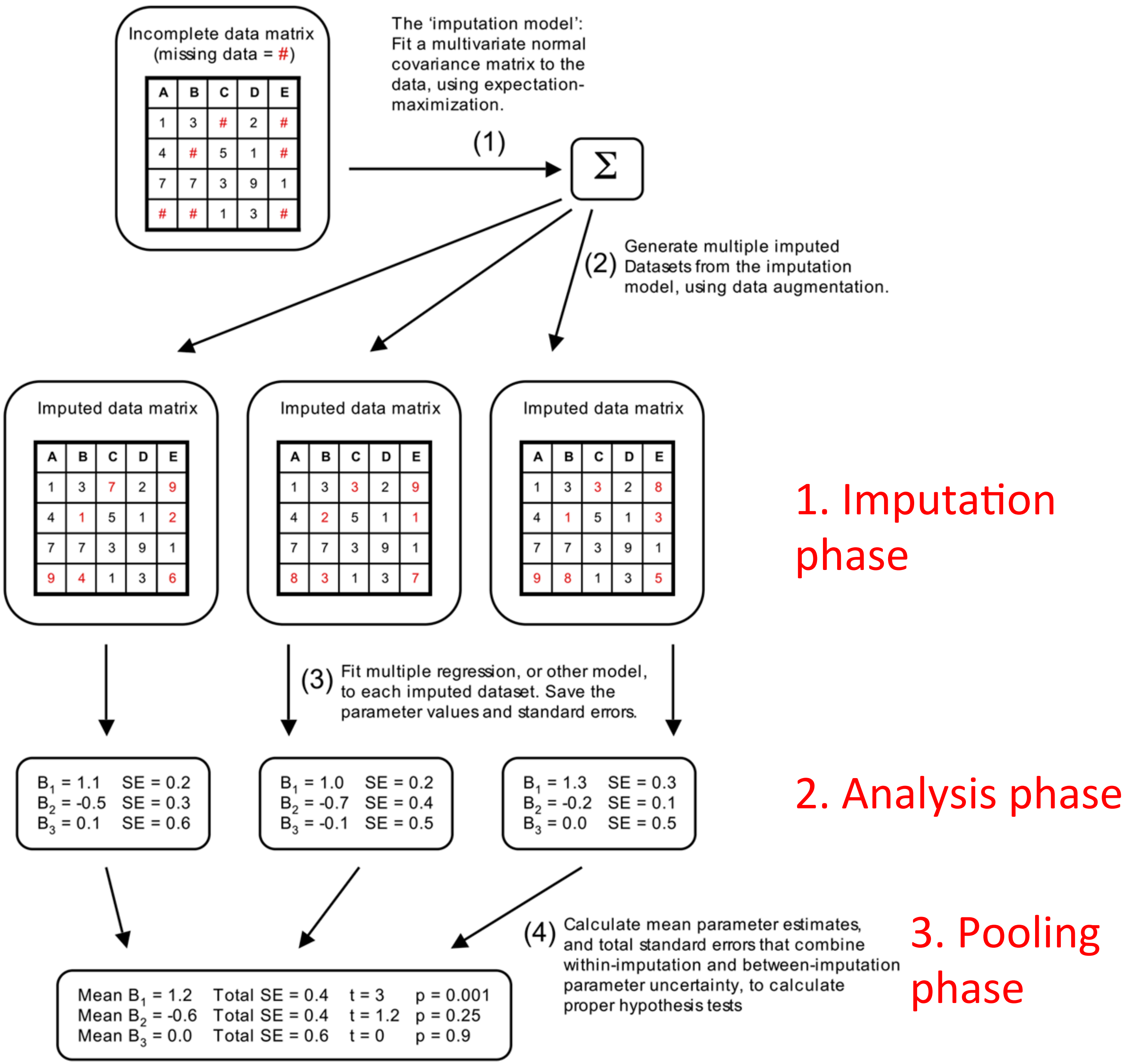


Example

x	x_{miss}
1	.
2	2
3	.
4	4
$\bar{x} = 2.5$	$\bar{x} = 3.0$

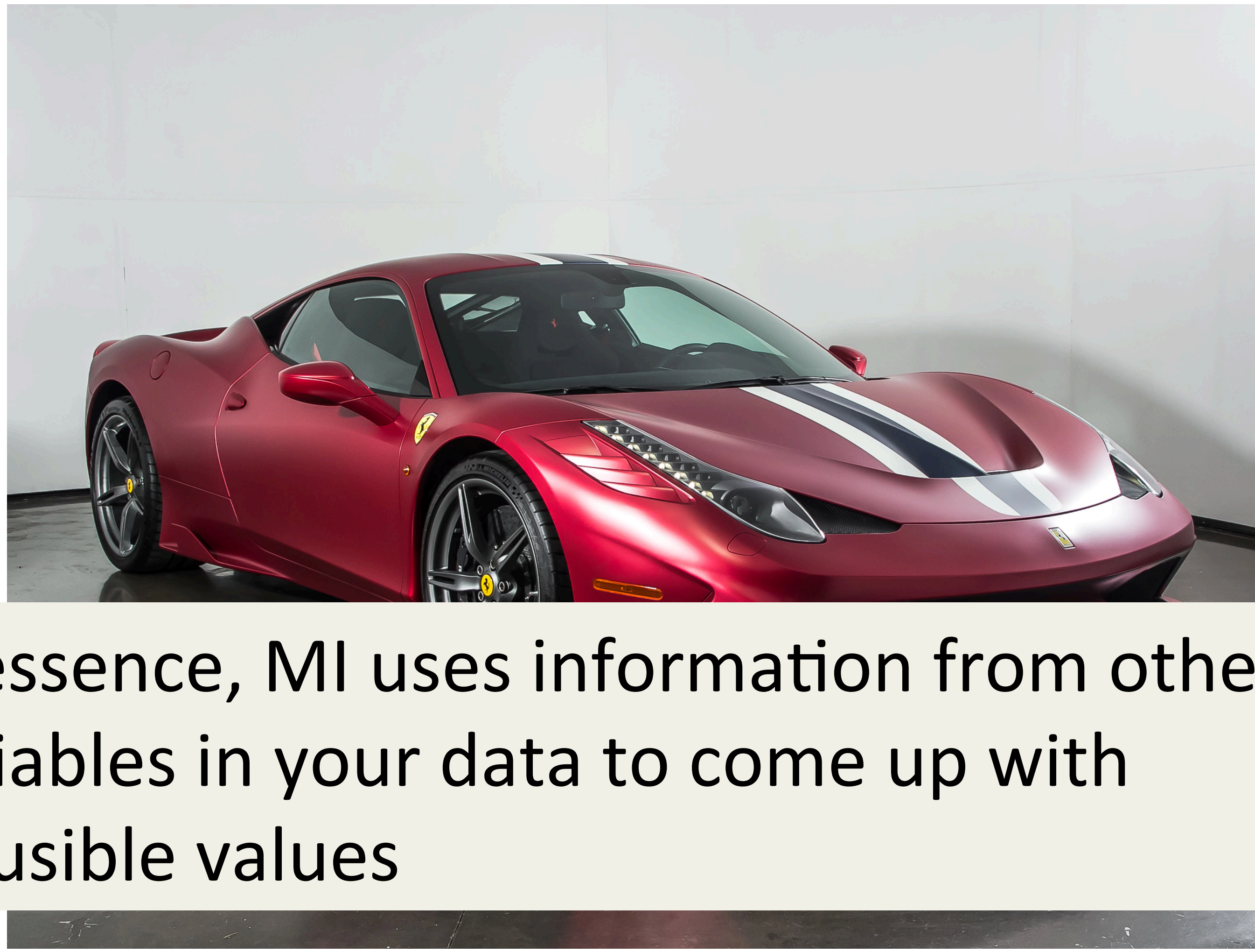
Example

x	x_{miss}	x_{MI1}	x_{MI2}
1	.	2	0
2	2	2	2
3	.	2	4
4	4	4	4
$\overline{x} = 2.5$	$\overline{x} = 3.0$	$\overline{x} = 2.5$	$\overline{x} = 2.5$



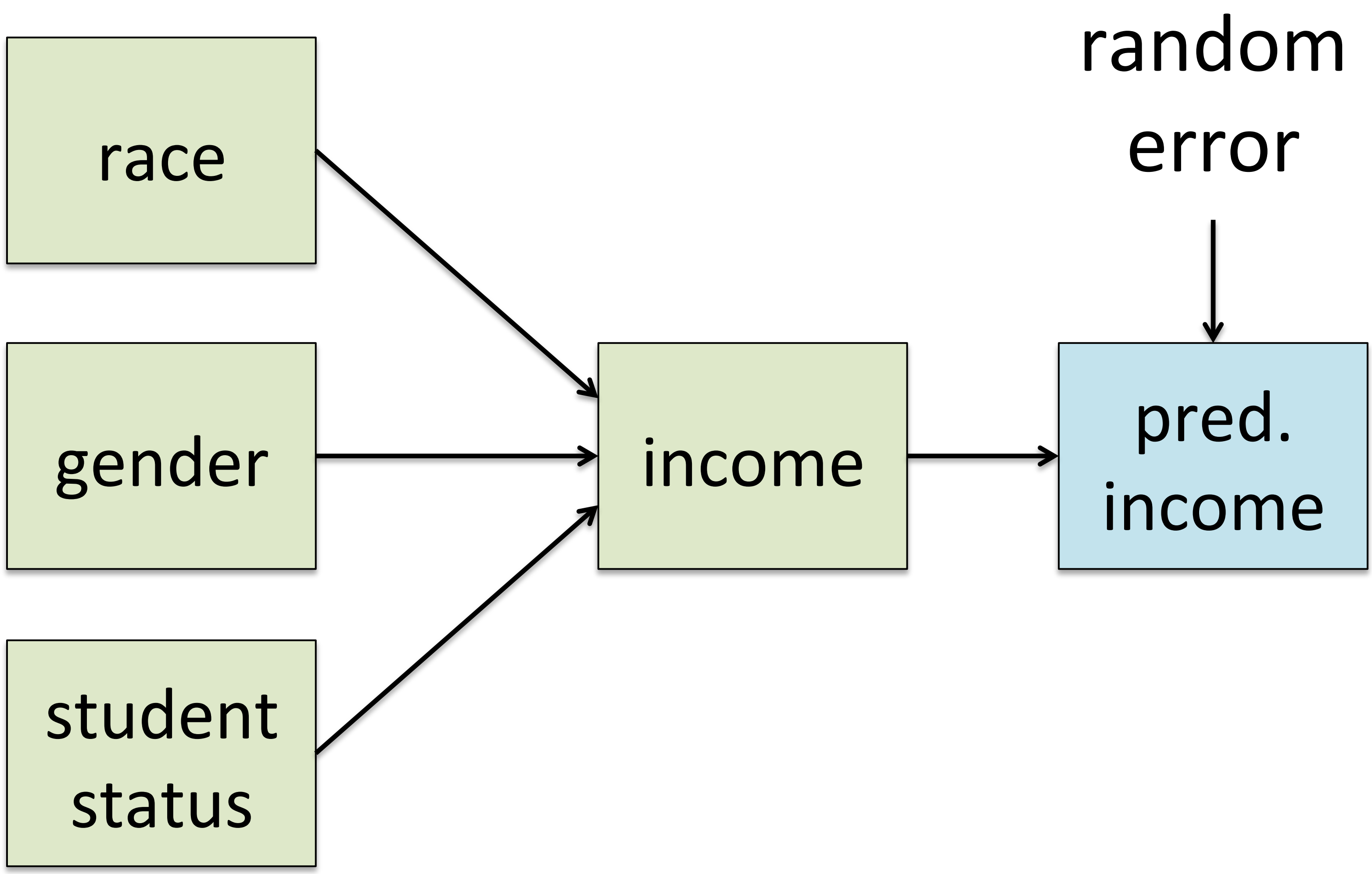
Incomplete data matrix
(missing data = #)
A B C D E
1 3 # 2 #
4 # 5 1 #
7 7 3 9 1
1 3
The 'imputation model':
Fit a multivariate normal covariance matrix to the data, using expectation-maximization.
(1)
Σ
(2) Generate multiple imputed Datasets from the imputation model, using data augmentation.
Imputed data matrix
A B C D E
1 3 7 2 9
4 1 5 1 2
7 7 3 9 1
9 4 1 3 6
Imputed data matrix
A B C D E
1 3 3 2 9
4 2 5 1 1
7 7 3 9 1
8 3 1 3 7
Imputed data matrix
A B C D E
1 3 3 2 8
4 1 5 1 3
7 7 3 9 1
9 8 1 3 5
1. Imputation phase
(3) Fit multiple regression, or other model, to each imputed dataset. Save the parameter values and standard errors.
B1 = 1.1 SE = 0.2
B2 = -0.5 SE = 0.3
B3 = 0.1 SE = 0.6
B1 = 1.0 SE = 0.2
B2 = -0.7 SE = 0.4
B3 = -0.1 SE = 0.5
B1 = 1.3 SE = 0.3
B2 = -0.2 SE = 0.1
B3 = 0.0 SE = 0.5
2. Analysis phase
(4) Calculate mean parameter estimates, and total standard errors that combine within-imputation and between-imputation parameter uncertainty, to calculate proper hypothesis tests
3. Pooling phase
Mean B1 = 1.2 Total SE = 0.4 t = 3 p = 0.001
Mean B2 = -0.6 Total SE = 0.4 t = 1.2 p = 0.25
Mean B3 = 0.0 Total SE = 0.6 t = 0 p = 0.9

Imputation phase: How MI makes up good values



In essence, MI uses information from other variables in your data to come up with plausible values

Imputation phase: How MI makes up good values



Why not use just one imputation?

- unbiased
- SE's too small
  - misses uncertainty due to missingness
- multiple imputed values reintroduces uncertainty due to missingness

Analysis Phase

imputed data 1	imputed data 2	imputed data 3
$\bar{y}_1 = \sum y_1 / n$	$\bar{y}_2 = \sum y_2 / n$	$\bar{y}_3 = \sum y_3 / n$
$\bar{y}_1 = 2$	$\bar{y}_2 = 2.1$	$\bar{y}_3 = 1.9$

Pooling Phase

take the mean of the *m* estimates

$\bar{y}_1 = 2$ $\bar{y}_2 = 2.1$ $\bar{y}_3 = 1.9$

$$\bar{y}_{pooled} = 2$$

Pooling Phase

works the same with regression coefficients

$\beta_{m=1} = 0.4$

$\beta_{m=2} = 0.3$

$\beta_{m=3} = 0.45$

$\beta_{pooled} = (0.4 + 0.3 + 0.45)/3$

$\beta_{pooled} = 0.38$

Pooling Phase: Standard Errors

- two sources of uncertainty
  - *within* imputations
  - *between* imputations

Within imputation variance

- Within imputation variance = mean of variances in each imputation

$$V_W = \frac{1}{m}\sum_{t=1}^{m} SE_t^2$$

Between imputation variance

- uncertainty due to missing data
- between imputation variance = variance of estimated statistics from the *m* analyses

$$V_B = \frac{1}{m-1}\sum_{t=1}^{m}\left(\hat{\beta}_t - \bar{\beta}\right)^2$$

average β across all models

β from estimation using data set t

Total Sampling Variance

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

adjusts for estimation using finite number of imputations

$$SE = \sqrt{V_T}$$

Exercise: Calculate β and SE_β by hand

- Below is a table with regression results in each of 5 imputed data sets – calculate the pooled MI estimates of β and SE_β

m	Variable	Estimate	S.E.
1	Income	.061	.022
2	Income	.033	.009
3	Income	.045	.012
4	Income	.071	.028
5	Income	.055	.015

$$V_W = \frac{1}{m}\sum_{t=1}^{m} SE_t^2$$

$$V_B = \frac{1}{m-1}\sum_{t=1}^{m}\left(\hat{\beta}_t - \bar{\beta}\right)^2$$

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

R CODE

```
beta=c(.061, .033, .045, .071, .055)
se=c(.022, .009, .012, .028, .015)

#MI point estimate
mean(beta)

#MI standard error
v.within=mean(se^2)
v.between=var(beta)
v.total=v.within+v.between+
(v.between/5)
sqrt(v.total)
```

RESULTS

β=.053

SE_{β}=.025

STATA CODE

```
input beta beta_se
.061 .022
.033 .009
.045 .012
.071 .028
.055 .015
end

/*MI point estimate*/
mean beta

/*MI standard error*/
gen se_sq = beta_se^2
su se_sq, d
scalar v_within = r(mean)
su beta, d
scalar v_between = r(Var)
scalar v_total = v_within + v_between +
v_between/5
di sqrt(v_total)
```

Two Major Approaches to MI

- Assume multivariate normal data (MVN)
  - often what people refer to when they talk about "multiple imputation"
- Allow data types to vary (ordinal, binary, etc.)
  - called "multiple imputation by chained equations", aka MICE
  - or fully conditional specification (FCS)

Comparing the two MI methods

- MVN has a solid theoretical basis
- MICE does not, but it has considerable intuitive appeal
- In practice…
  - both tend to give comparable results*
  - MVN tends to be faster
  - but MVN might not converge with many non-continuous/normal variables

MVN imputation

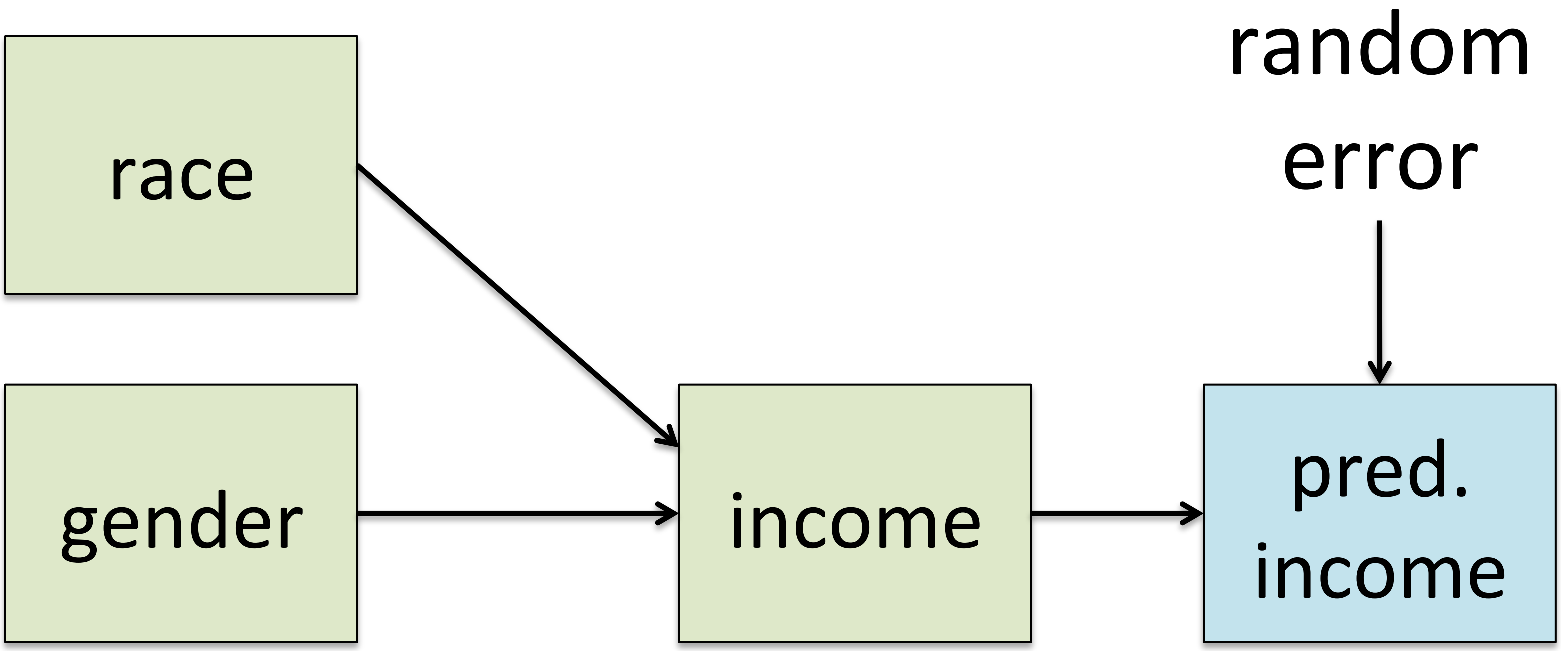
- today we will focus on MVN imputation
- focus on data augmentation

	Stata	R	SPSS
MVN	mi mvn	Amelia; norm	N/A
MICE	mi chained	mi; mice	multiple imputation

MVN imputation – data augmentation

$$E(income) = \beta_0 + \beta_1 * white + \beta_2 * male$$

$$income_i = \hat{y} + z_i$$



MVN imputation – data augmentation

$$\mu_1 \Sigma_1 \rightarrow \mu_1^* \Sigma_1^*$$

$$E(income) = \beta_0 + \beta_1 * white + \beta_2 * male$$

$$income_i = \hat{y} + z_i$$

$$\mu_2 \Sigma_2 \rightarrow \mu_2^* \Sigma_2^*$$

Data augmentation example

20% of test scores missing

$$testscore_i^* = \beta_0 + \beta_1 health_i + z_i$$

predicted value	random residual	imputed value
50.37	-28.57	21.79
50.48	27.52	78.00
50.34	-11.68	38.67
50.14	13.99	64.14
50.77	-12.51	38.26
50.13	-5.10	45.03
50.36	5.53	55.89
50.11	2.19	52.30
50.26	-20.04	30.22

	Means		**Variances**		**Covariance**
	Health	Test score	Health	Test score	
Complete data	55.4	49.5	883.8	355.1	104.1
iteration 1	55.4	51.2	883.8	353.9	28.6

Data augmentation example

	Means		Variances		Covariance
	Health	Test score	Health	Test score	
Complete data	55.4	49.5	883.8	355.1	104.1
iteration 1	55.4	51.2	883.8	353.9	28.6

	Means		Variances		Covariance
	Health	Test score	Health	Test score	
iteration 1	55.4	51.2	883.8	353.9	28.6
noise added	62.1	38.9	922.2	319.8	45.5

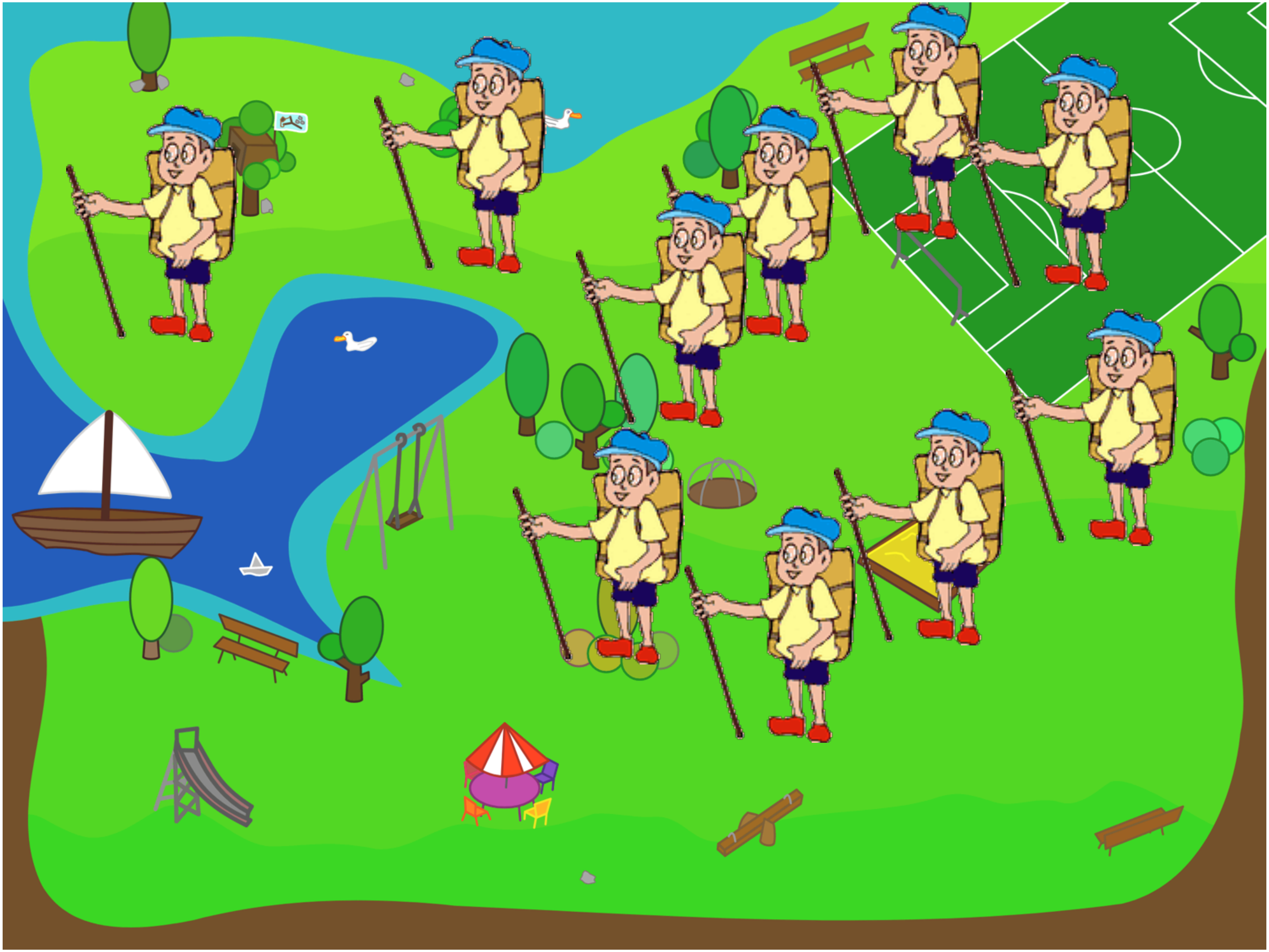
$$testscore_i^* = \beta_0 + \beta_1 health_i + z_i$$

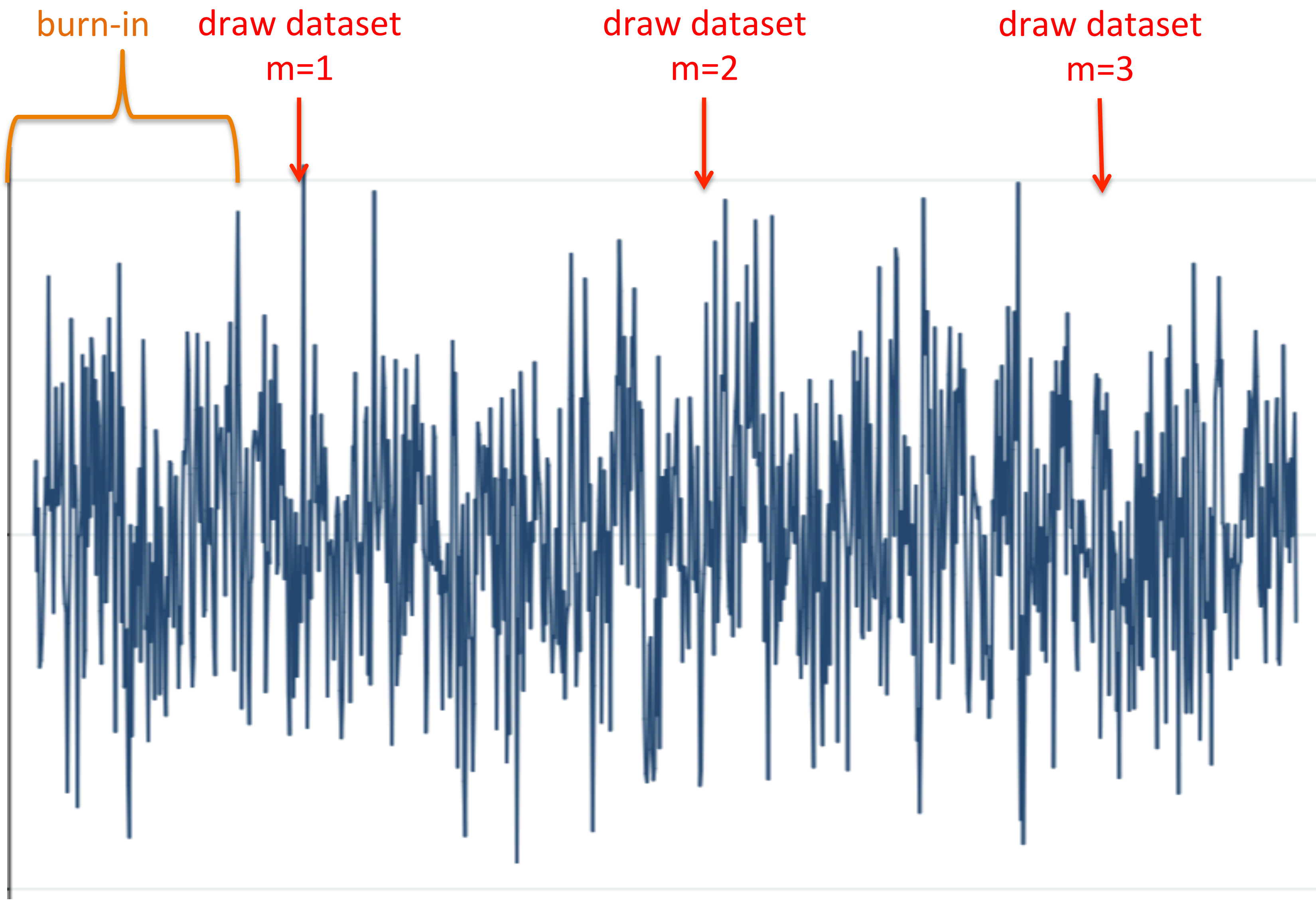
Data augmentation example

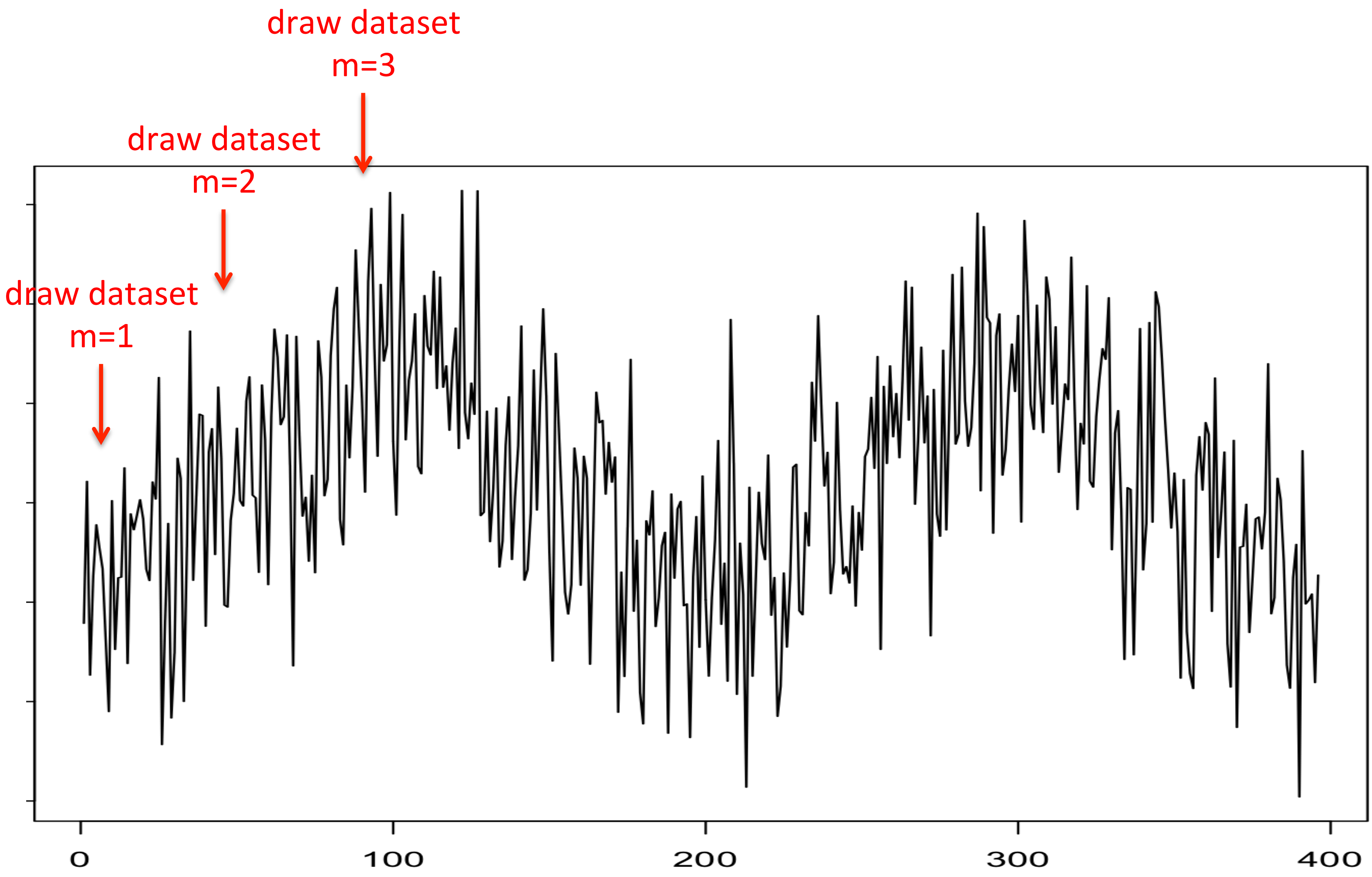
	Means		Variances		Covariance
	Health	Test score	Health	Test score	
Complete data	55.4	49.5	883.8	355.1	104.1
iteration 1	55.4	51.2	883.8	353.9	28.6
iteration 2	55.4	53.0	883.8	378.1	-50.5
iteration 3	55.4	52.2	883.8	354.2	-22.6
iteration 4	55.4	50.1	883.8	336.9	64.4
iteration 5	55.4	51.3	883.8	343.7	12.2

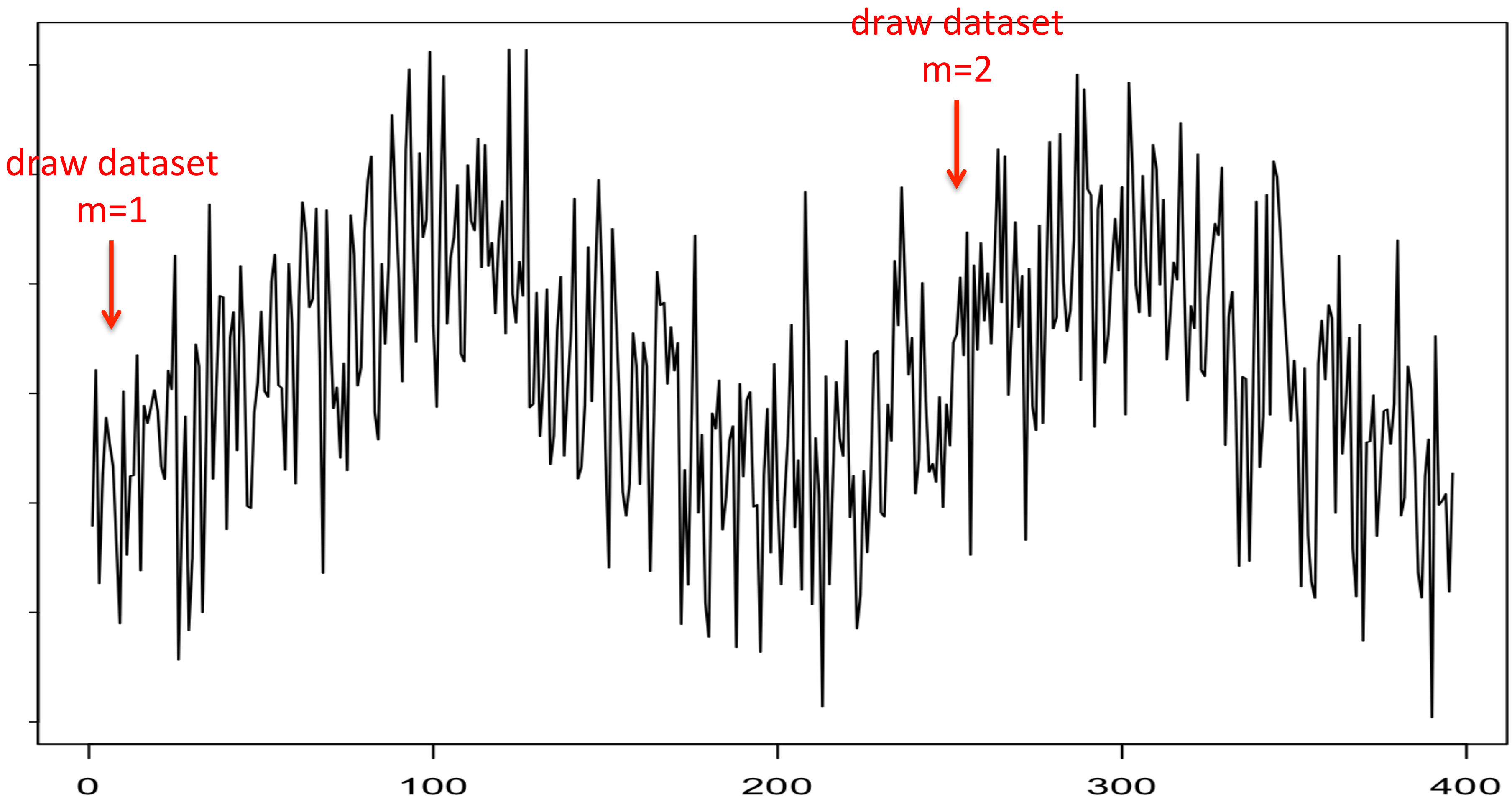
MVN imputation – data augmentation

- constant stream of parameters
- we want to sample from all over the parameter space
- close iterations likely to be correlated
- let model run in between taking imputed data sets

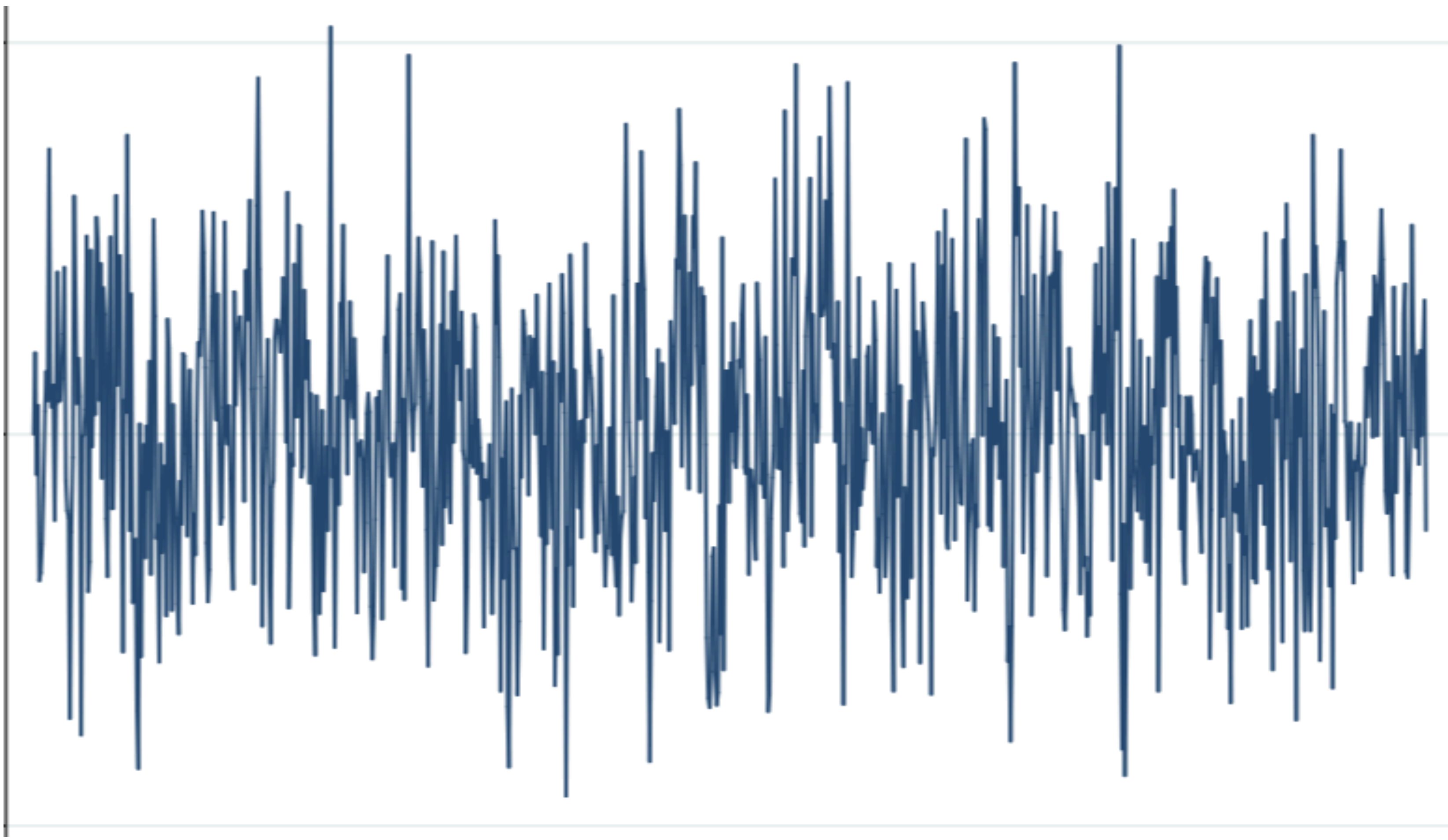



burn-in
draw dataset
m=1
draw dataset
m=2
draw dataset
m=3


draw dataset
m=3
draw dataset
m=2
draw dataset
m=1
0
100
200
300
400

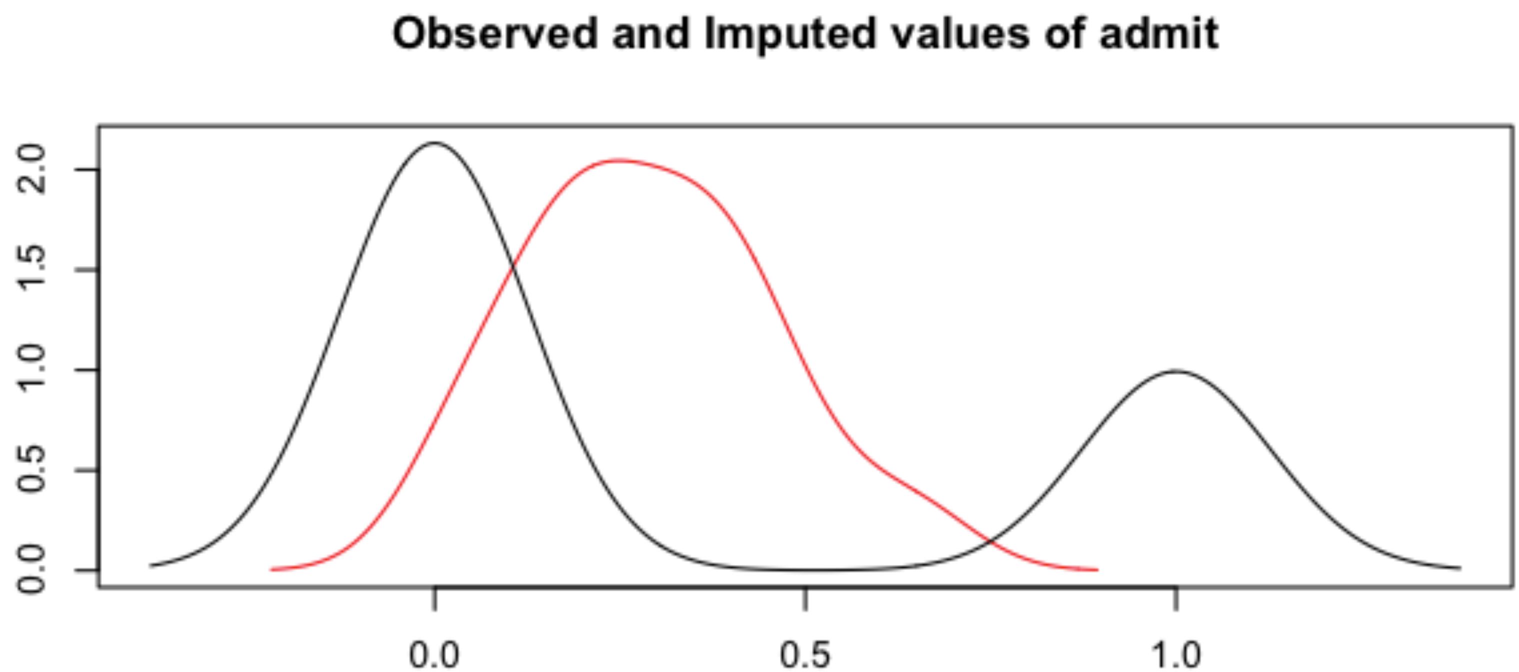

draw dataset
m=1
draw dataset
m=2
0
100
200
300
400

Worst Linear Function



Implications of using MVN imputation

- multivariate normal imputation model can impute strange values, e.g., binary variables with imputed values of 0.3



Chained Equations

- predict each variable with most appropriate type of regression

$\text{logit(married)} = \beta_0 + \beta_1\text{age} + \beta_2\text{race} + \beta_3\text{religion}$

$\text{poisson(children)} = \beta_0 + \beta_1\text{age} + \beta_2\text{race} + \beta_3\text{married}$

Rules for MI

- Include in your imputation equations any variables that:
  - will be used in your final analysis (including the outcome)
  - any variables that predict missingness
  - any variables that are highly correlated with the variables you want to impute (i.e., have lots of information for making good imputations)
- also include any higher order terms that might be of interest (e.g., interactions, squares)
  - failure to do so can bias results towards 0

How many imputations?

- It depends on how much missing information there is

$$FMI = \frac{V_B + V_B / m}{V_T}$$

How many imputations?

- more imputations means more statistical power

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

- more imputations makes your results more reproducible

rule of thumb – at least as many imputations as the percentage of cases with missing data

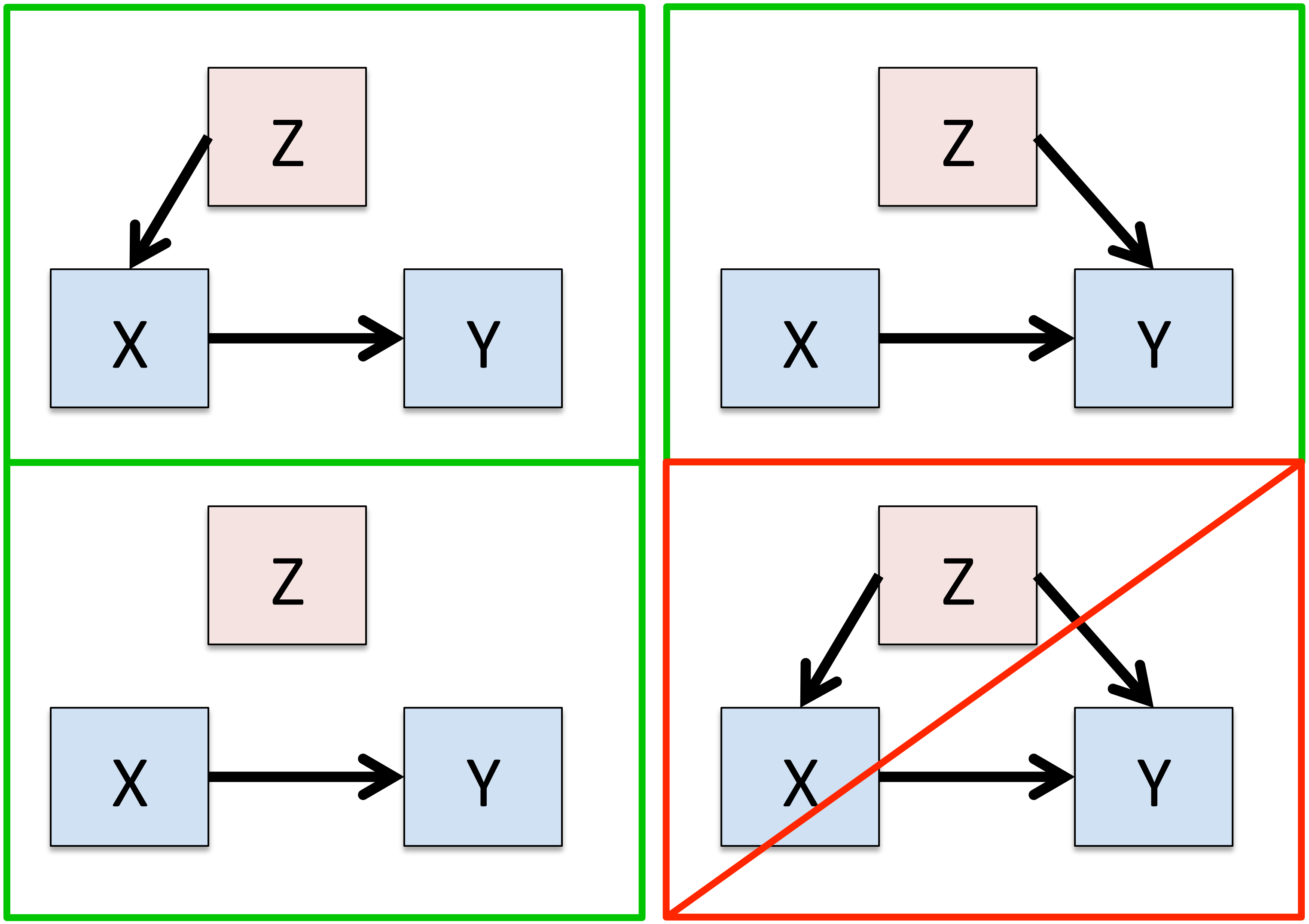
What you lose using MI

- In general, “statistics whose value changes systematically with the sample size cannot be combined using Rubin’s rules”*
  - e.g., AIC, BIC, likelihood ratio test
- Time

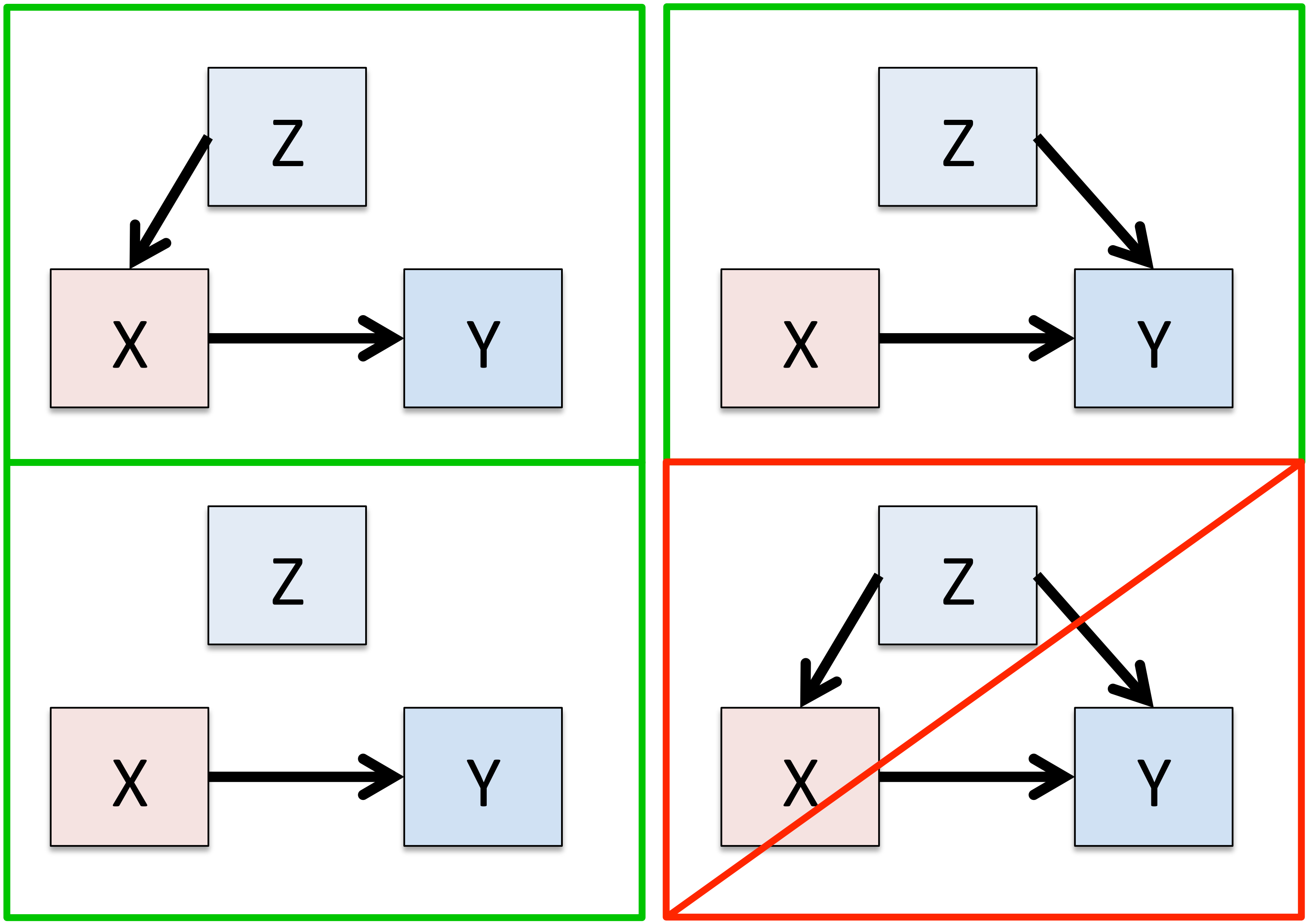
When to Use Multiple Imputation

- Maximize efficiency with MCAR or MAR data
- Descriptive statistics
- Regression – missingness depends on Y

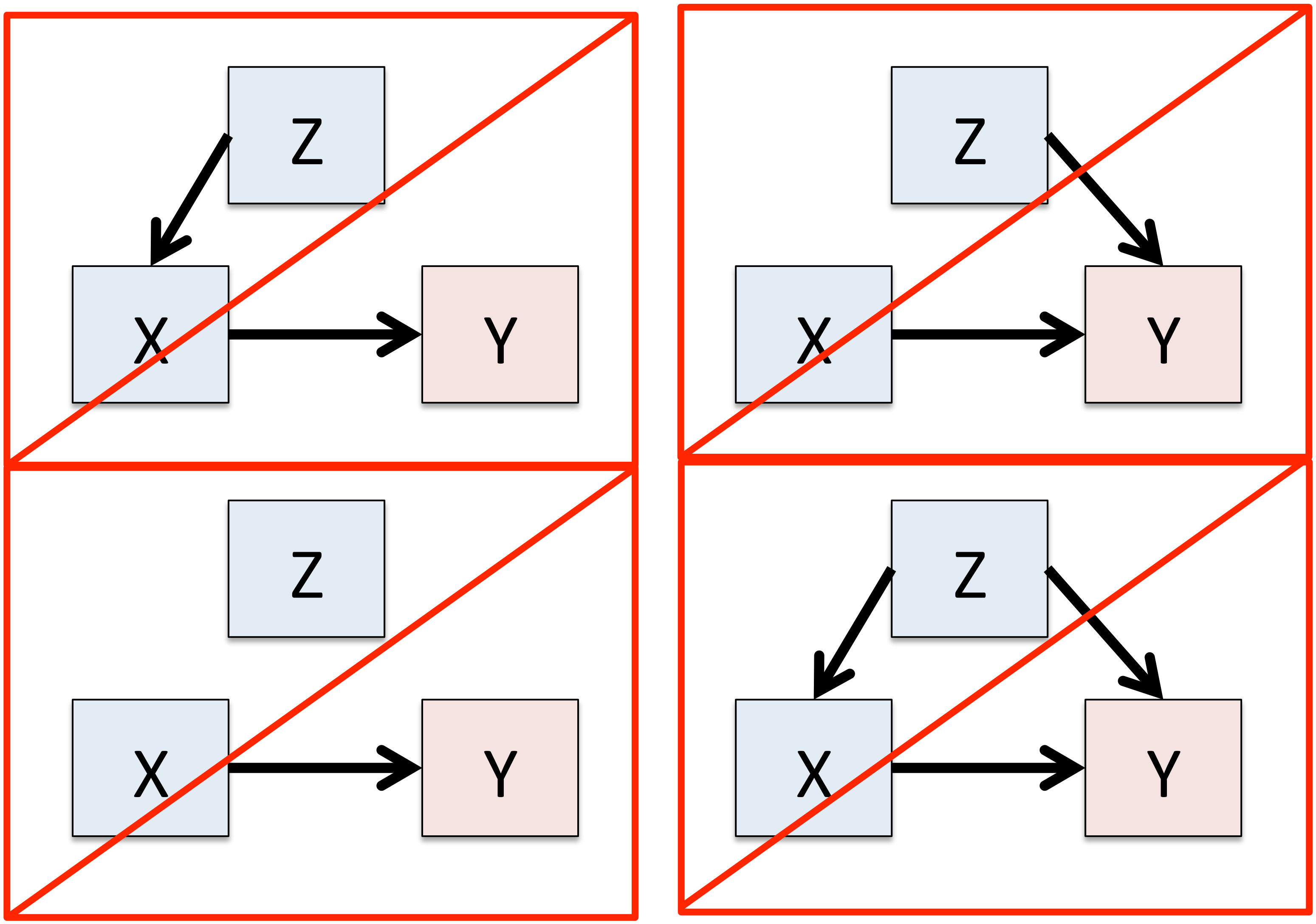
When is Listwise Deletion Unbiased?



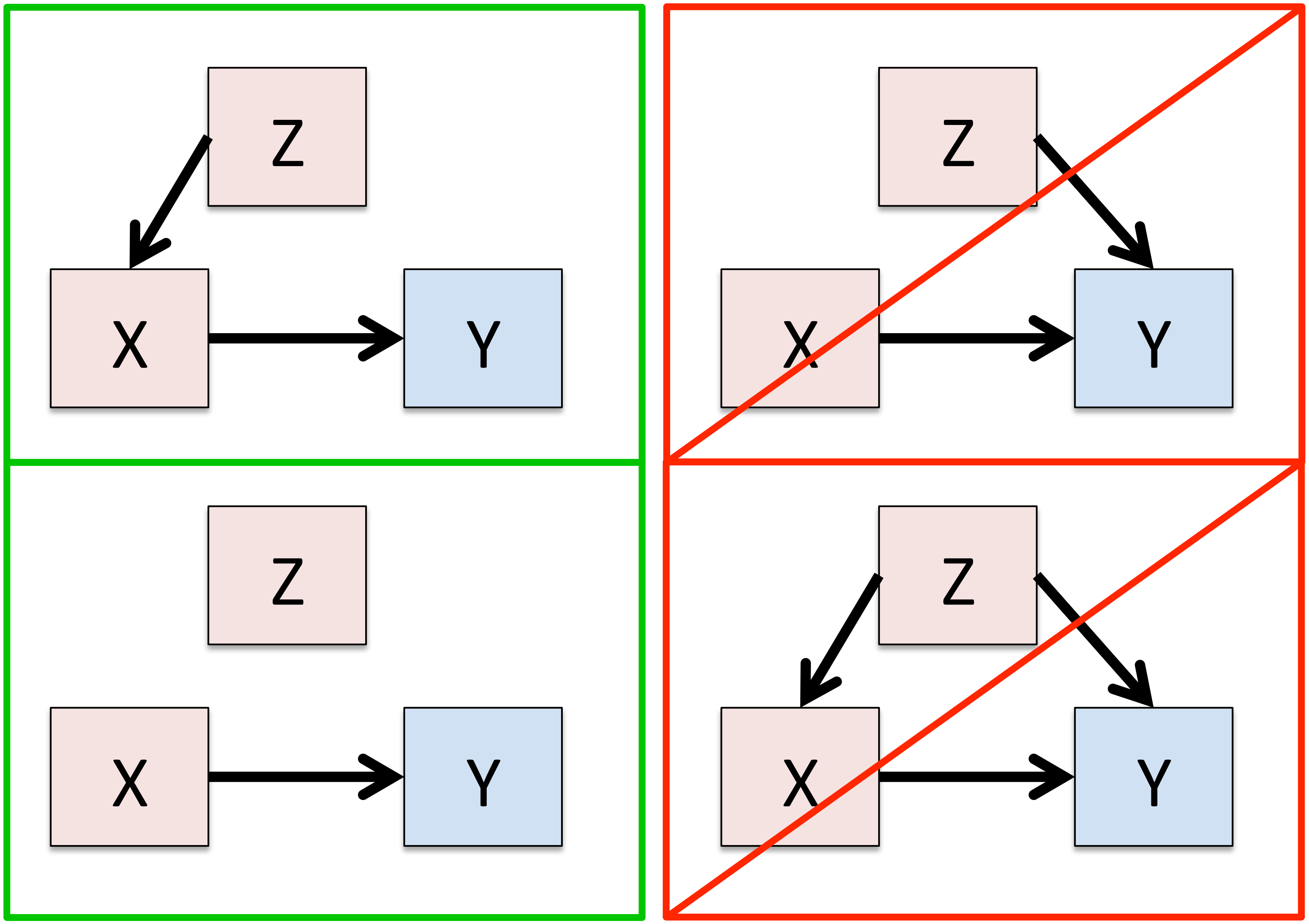
When is Listwise Deletion Unbiased?



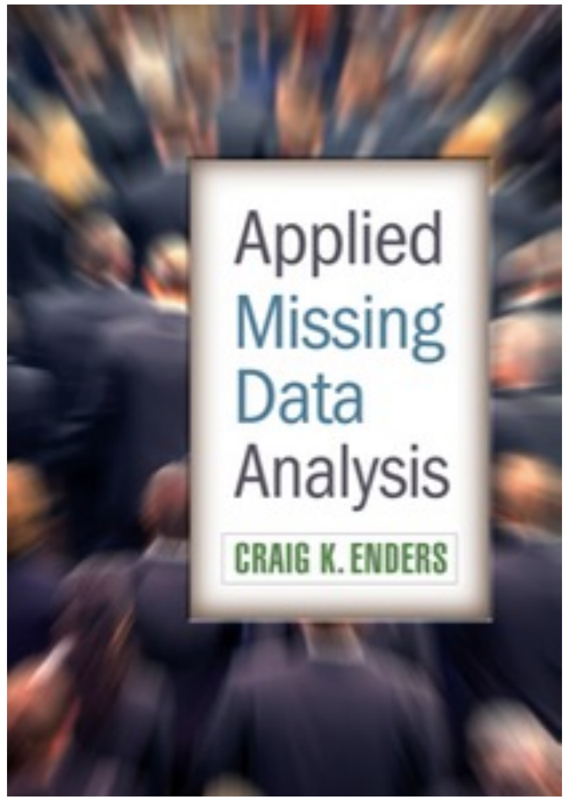
When is Listwise Deletion Unbiased?



When is Listwise Deletion Unbiased?



Further Reading



References

- https://pictures.dealer.com/b/boardwalkferrari/1685/4ef1f5b56b86488255a6c45e8be2ed9bx.jpg
- https://upload.wikimedia.org/wikipedia/commons/thumb/5/5c/Stata_Logo.svg/2000px-Stata_Logo.svg.png
- http://worldartsme.com/images/cartoon-hiker-clipart-1.jpg
- http://www.clker.com/cliparts/2/f/3/9/134557324648652724Park%20by%20the%20Sea.svg
- http://www.animatedimages.org/img-animated-hiking-image-0009-173652.htm