Quantitative Social Science Methods, I, Lecture Notes: Missing Data

Gary King[1]
Institute for Quantitative Social Science
Harvard University

August 17, 2020

[1] GaryKing.org

Summary

Summary

- Biased or inefficient missing data practices

Summary

- Biased or inefficient missing data practices
  - Make up numbers: e.g., change Party ID “don’t knows” to “independent”

Summary

- Biased or inefficient missing data practices
  - Make up numbers: e.g., change Party ID "don't knows" to "independent"
  - Listwise deletion: used by 94% pre-2000 in AJPS/APSR/BJPS

Summary

- Biased or inefficient missing data practices
  - Make up numbers: e.g., change Party ID “don’t knows” to “independent”
  - Listwise deletion: used by 94% pre-2000 in AJPS/APSR/BJPS
  - Many other ad hoc approaches: some intuitive, most biased

Summary

- Biased or inefficient missing data practices
  - Make up numbers: e.g., change Party ID "don't knows" to "independent"
  - Listwise deletion: used by 94% pre-2000 in AJPS/APSR/BJPS
  - Many other ad hoc approaches: some intuitive, most biased
- Application-specific methods: efficient, but model-dependent, hard to develop and use

Summary

- Biased or inefficient missing data practices
  - Make up numbers: e.g., change Party ID “don’t knows” to “independent”
  - Listwise deletion: used by 94% pre-2000 in AJPS/APSR/BJPS
  - Many other ad hoc approaches: some intuitive, most biased
- Application-specific methods: efficient, but model-dependent, hard to develop and use
- Multiple imputation: An easy-to-use, statistically appropriate alternative

Summary

- Biased or inefficient missing data practices
  - Make up numbers: e.g., change Party ID "don't knows" to "independent"
  - Listwise deletion: used by 94% pre-2000 in AJPS/APSR/BJPS
  - Many other ad hoc approaches: some intuitive, most biased
- Application-specific methods: efficient, but model-dependent, hard to develop and use
- Multiple imputation: An easy-to-use, statistically appropriate alternative
  - fill in ≈ 5 data sets with different imputations for missing values

Summary

- Biased or inefficient missing data practices
  - Make up numbers: e.g., change Party ID “don’t knows” to “independent”
  - Listwise deletion: used by 94% pre-2000 in AJPS/APSR/BJPS
  - Many other ad hoc approaches: some intuitive, most biased
- Application-specific methods: efficient, but model-dependent, hard to develop and use
- Multiple imputation: An easy-to-use, statistically appropriate alternative
  - fill in ≈ 5 data sets with different imputations for missing values
  - Convenience: analyze each one as you would without missingness

Summary

- Biased or inefficient missing data practices
  - Make up numbers: e.g., change Party ID “don’t knows” to “independent”
  - Listwise deletion: used by 94% pre-2000 in AJPS/APSR/BJPS
  - Many other ad hoc approaches: some intuitive, most biased
- Application-specific methods: efficient, but model-dependent, hard to develop and use
- Multiple imputation: An easy-to-use, statistically appropriate alternative
  - fill in ≈ 5 data sets with different imputations for missing values
  - Convenience: analyze each one as you would without missingness
  - Use a special method to combine the results

Summary

- Biased or inefficient missing data practices
  - Make up numbers: e.g., change Party ID "don't knows" to "independent"
  - Listwise deletion: used by 94% pre-2000 in AJPS/APSR/BJPS
  - Many other ad hoc approaches: some intuitive, most biased
- Application-specific methods: efficient, but model-dependent, hard to develop and use
- Multiple imputation: An easy-to-use, statistically appropriate alternative
  - fill in ≈ 5 data sets with different imputations for missing values
  - Convenience: analyze each one as you would without missingness
  - Use a special method to combine the results
  - Robust: separate missingness and analysis models

Readings

`j.mp/MisMeas`

Readings

- Gary King; James Honaker; Anne ~~Joseph~~ O'Connell; Kenneth Scheve. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation" APSR, 2001.

`j.mp/MisMeas`

Readings

- Gary King; James Honaker; Anne ~~Joseph~~ O'Connell; Kenneth Scheve. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation" APSR, 2001.
- James Honaker and Gary King. "What to do about Missing Values in Time Series Cross-Section Data" AJPS, 2010.

j.mp/MisMeas

Readings

- Gary King; James Honaker; Anne ~~Joseph~~ O'Connell; Kenneth Scheve. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation" APSR, 2001.
- James Honaker and Gary King. "What to do about Missing Values in Time Series Cross-Section Data" AJPS, 2010.
- Blackwell, Matthew, James Honaker, and Gary King. "A Unified Approach to Measurement Error and Missing Data: {Overview, Details and Extensions}" SMR, 2017.

j.mp/MisMeas

Readings

- Gary King; James Honaker; Anne ~~Joseph~~ O'Connell; Kenneth Scheve. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation" APSR, 2001.
- James Honaker and Gary King. "What to do about Missing Values in Time Series Cross-Section Data" AJPS, 2010.
- Blackwell, Matthew, James Honaker, and Gary King. "A Unified Approach to Measurement Error and Missing Data: {Overview, Details and Extensions}" SMR, 2017.
- Amelia II: A Program for Missing Data

j.mp/MisMeas

Notation

Notation

$$D = \begin{pmatrix} 1 & 2.5 & 432 & 0 \\ 5 & 3.2 & 543 & 1 \\ 2 & 7.4 & 219 & 1 \\ 6 & 1.9 & 234 & 1 \\ 3 & 1.2 & 108 & 0 \\ 0 & 7.7 & 95 & 1 \end{pmatrix}, \qquad M = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

Notation

$$D = \begin{pmatrix} 1 & 2.5 & 432 & 0 \\ 5 & 3.2 & 543 & 1 \\ 2 & 7.4 & 219 & 1 \\ 6 & 1.9 & 234 & 1 \\ 3 & 1.2 & 108 & 0 \\ 0 & 7.7 & 95 & 1 \end{pmatrix}, \qquad M = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

D_{mis} = missing elements in D

Notation

$$D = \begin{pmatrix} 1 & 2.5 & 432 & 0 \\ 5 & 3.2 & 543 & 1 \\ 2 & 7.4 & 219 & 1 \\ 6 & 1.9 & 234 & 1 \\ 3 & 1.2 & 108 & 0 \\ 0 & 7.7 & 95 & 1 \end{pmatrix}, \qquad M = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

D_{mis} = missing elements in D
D_{obs} = observed elements in D

Notation

$$D = \begin{pmatrix} 1 & 2.5 & 432 & 0 \\ 5 & 3.2 & 543 & 1 \\ 2 & 7.4 & 219 & 1 \\ 6 & 1.9 & 234 & 1 \\ 3 & 1.2 & 108 & 0 \\ 0 & 7.7 & 95 & 1 \end{pmatrix}, \qquad M = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

D_{mis} = missing elements in D
D_{obs} = observed elements in D

- If you treat some elements as missing: make sure they exist

Notation

$$D = \begin{pmatrix} 1 & 2.5 & 432 & 0 \\ 5 & 3.2 & 543 & 1 \\ 2 & 7.4 & 219 & 1 \\ 6 & 1.9 & 234 & 1 \\ 3 & 1.2 & 108 & 0 \\ 0 & 7.7 & 95 & 1 \end{pmatrix}, \qquad M = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

D_{mis} = missing elements in D
D_{obs} = observed elements in D

- If you treat some elements as missing: make sure they exist (what's your view on the National Helium Reserve?)

Notation

$$D = \begin{pmatrix} 1 & 2.5 & 432 & 0 \\ 5 & 3.2 & 543 & 1 \\ 2 & \color{blue}{7.4} & 219 & \color{blue}{1} \\ \color{blue}{6} & 1.9 & \color{blue}{234} & 1 \\ \color{blue}{3} & 1.2 & 108 & 0 \\ \color{blue}{0} & 7.7 & 95 & 1 \end{pmatrix}, \qquad M = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

D_{mis} = missing elements in D
D_{obs} = observed elements in D

- If you treat some elements as missing: make sure they exist (what's your view on the National Helium Reserve?)
- Quiz: If some values don't exist, what analysis would you do?

Missingness Assumptions

Missingness Assumptions

Please excuse the literature's crazy nomenclature

Missingness Assumptions

Please excuse the literature's crazy nomenclature

Assumption	Acronym	Can predict M with:
Missing Completely At Random	MCAR	—

Missingness Assumptions

Please excuse the literature's crazy nomenclature

Assumption	Acronym	Can predict M with:
Missing Completely At Random	MCAR	—

Missingness Assumptions

Please excuse the literature's crazy nomenclature

Assumption	Acronym	Can predict M with:
Missing Completely At Random	MCAR	—
Missing At Random	MAR	D_{obs}

Missingness Assumptions

Please excuse the literature's crazy nomenclature

Assumption	Acronym	Can predict M with:
Missing Completely At Random	MCAR	—
Missing At Random	MAR	D_{obs}
Nonignorable	NI	D_{obs} & D_{mis}

Missingness Assumptions, again

Missingness Assumptions, again

1. MCAR (naive): Coin flips determine answering survey questions

Missingness Assumptions, again

1. MCAR (naive): Coin flips determine answering survey questions

$$P(M|D) = P(M)$$

Missingness Assumptions, again

1. MCAR (naive): Coin flips determine answering survey questions
$$P(M|D) = P(M)$$

2. MAR (empirical): missingness is a function of measured variables

Missingness Assumptions, again

1. MCAR (naive): Coin flips determine answering survey questions

$$P(M|D) = P(M)$$

2. MAR (empirical): missingness is a function of measured variables

$$P(M|D) \equiv P(M|D_{obs}, D_{mis}) = P(M|D_{obs})$$

Missingness Assumptions, again

1. MCAR (naive): Coin flips determine answering survey questions

$$P(M|D) = P(M)$$

2. MAR (empirical): missingness is a function of measured variables

$$P(M|D) \equiv P(M|D_{obs}, D_{mis}) = P(M|D_{obs})$$

   - e.g., Independents are less likely to answer vote choice question (with PID measured)

Missingness Assumptions, again

1. MCAR (naive): Coin flips determine answering survey questions

$$P(M|D) = P(M)$$

2. MAR (empirical): missingness is a function of measured variables

$$P(M|D) \equiv P(M|D_{obs}, D_{mis}) = P(M|D_{obs})$$

- e.g., Independents are less likely to answer vote choice question (with PID measured)
- e.g., Some occupations are less likely to answer the income question (with occupation measured)

Missingness Assumptions, again

1. MCAR (naive): Coin flips determine answering survey questions

$$P(M|D) = P(M)$$

2. MAR (empirical): missingness is a function of measured variables

$$P(M|D) \equiv P(M|D_{obs}, D_{mis}) = P(M|D_{obs})$$

   - e.g., Independents are less likely to answer vote choice question (with PID measured)
   - e.g., Some occupations are less likely to answer the income question (with occupation measured)

3. NI (fatalistic): missingness depends on unobservables

Missingness Assumptions, again

1. MCAR (naive): Coin flips determine answering survey questions
$$P(M|D) = P(M)$$
2. MAR (empirical): missingness is a function of measured variables
$$P(M|D) \equiv P(M|D_{obs}, D_{mis}) = P(M|D_{obs})$$
   - e.g., Independents are less likely to answer vote choice question (with PID measured)
   - e.g., Some occupations are less likely to answer the income question (with occupation measured)
3. NI (fatalistic): missingness depends on unobservables
   - $P(M|D)$ doesn't simplify

Missingness Assumptions, again

1. MCAR (naive): Coin flips determine answering survey questions

$$P(M|D) = P(M)$$

2. MAR (empirical): missingness is a function of measured variables

$$P(M|D) \equiv P(M|D_{obs}, D_{mis}) = P(M|D_{obs})$$

   - e.g., Independents are less likely to answer vote choice question (with PID measured)
   - e.g., Some occupations are less likely to answer the income question (with occupation measured)

3. NI (fatalistic): missingness depends on unobservables
   - $P(M|D)$ doesn't simplify
   - e.g., censoring income if income is > $\$100K$ <u>and</u> you can't predict high income with other measured variables

Missingness Assumptions, again

1. MCAR (naive): Coin flips determine answering survey questions

$$P(M|D) = P(M)$$

2. MAR (empirical): missingness is a function of measured variables

$$P(M|D) \equiv P(M|D_{obs}, D_{mis}) = P(M|D_{obs})$$

   - e.g., Independents are less likely to answer vote choice question (with PID measured)
   - e.g., Some occupations are less likely to answer the income question (with occupation measured)

3. NI (fatalistic): missingness depends on unobservables
   - $P(M|D)$ doesn't simplify
   - e.g., censoring income if income is > $\$100K$ and you can't predict high income with other measured variables
   - Adding variables to predict income can change NI to MAR

Existing General Purpose Missing Data Methods

Existing General Purpose Missing Data Methods

- Listwise deletion. MCAR: RMSE is 1 SE off; MAR: biased

Existing General Purpose Missing Data Methods

- Listwise deletion. MCAR: RMSE is 1 SE off; MAR: biased
- Best guess imputation: Depends on guesser!

Existing General Purpose Missing Data Methods

- Listwise deletion. MCAR: RMSE is 1 SE off; MAR: biased
- Best guess imputation: Depends on guesser!
- Impute 0, add indicator to control: biased

Existing General Purpose Missing Data Methods

- Listwise deletion. MCAR: RMSE is 1 SE off; MAR: biased
- Best guess imputation: Depends on guesser!
- Impute 0, add indicator to control: biased
- Pairwise deletion: biased unless MCAR

Existing General Purpose Missing Data Methods

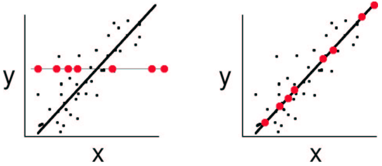
- Listwise deletion. MCAR: RMSE is 1 SE off; MAR: biased
- Best guess imputation: Depends on guesser!
- Impute 0, add indicator to control: biased
- Pairwise deletion: biased unless MCAR
- Hot deck imputation: Inefficient, SEs wrong

Existing General Purpose Missing Data Methods

- Listwise deletion. MCAR: RMSE is 1 SE off; MAR: biased
- Best guess imputation: Depends on guesser!
- Impute 0, add indicator to control: biased
- Pairwise deletion: biased unless MCAR
- Hot deck imputation: Inefficient, SEs wrong
- Mean substitution: attenuation bias

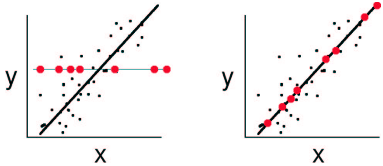
Existing General Purpose Missing Data Methods

- Listwise deletion. MCAR: RMSE is 1 SE off; MAR: biased
- Best guess imputation: Depends on guesser!
- Impute 0, add indicator to control: biased
- Pairwise deletion: biased unless MCAR
- Hot deck imputation: Inefficient, SEs wrong
- Mean substitution: attenuation bias



Existing General Purpose Missing Data Methods

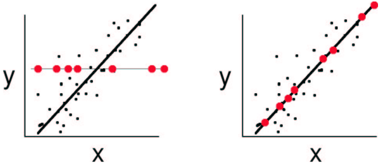
- Listwise deletion. MCAR: RMSE is 1 SE off; MAR: biased
- Best guess imputation: Depends on guesser!
- Impute 0, add indicator to control: biased
- Pairwise deletion: biased unless MCAR
- Hot deck imputation: Inefficient, SEs wrong
- Mean substitution: attenuation bias



- y-hat Regression Imputation: Optimistic (scatter when observed, perfectly linear when unobserved?); SEs too small

Existing General Purpose Missing Data Methods

- Listwise deletion. MCAR: RMSE is 1 SE off; MAR: biased
- Best guess imputation: Depends on guesser!
- Impute 0, add indicator to control: biased
- Pairwise deletion: biased unless MCAR
- Hot deck imputation: Inefficient, SEs wrong
- Mean substitution: attenuation bias



- y-hat Regression Imputation: Optimistic (scatter when observed, perfectly linear when unobserved?); SEs too small
- y-hat + ϵ Regression Imputation: Ignores estimation uncertainty; impossible with scattered missingness; getting there

Application-specific Methods: Overview

Application-specific Methods: Overview

- Examples: Models of censoring, truncation, etc.

Application-specific Methods: Overview

- Examples: Models of censoring, truncation, etc.
- Likelihood Inference: Unknowns given knowns, $P(\theta|Y_{obs})$.

Application-specific Methods: Overview

- Examples: Models of censoring, truncation, etc.
- Likelihood Inference: Unknowns given knowns, $P(\theta|Y_{obs})$.
- Optimal, if model is correct

Application-specific Methods: Overview

- Examples: Models of censoring, truncation, etc.
- Likelihood Inference: Unknowns given knowns, $P(\theta|Y_{obs})$.
- Optimal, if model is correct
- Can be model dependent: sensitive to distributional assumptions

Application-specific Methods: Overview

- Examples: Models of censoring, truncation, etc.
- Likelihood Inference: Unknowns given knowns, $P(\theta|Y_{obs})$.
- Optimal, if model is correct
- Can be model dependent: sensitive to distributional assumptions
- Often difficult practically: Unless code already exists

Application-specific Methods: Overview

- Examples: Models of censoring, truncation, etc.
- Likelihood Inference: Unknowns given knowns, $P(\theta|Y_{obs})$.
- Optimal, if model is correct
- Can be model dependent: sensitive to distributional assumptions
- Often difficult practically: Unless code already exists
- Rare with scattered missingness in X and Y

Application-specific Methods: Overview

- Examples: Models of censoring, truncation, etc.
- Likelihood Inference: Unknowns given knowns, $P(\theta|Y_{obs})$.
- Optimal, if model is correct
- Can be model dependent: sensitive to distributional assumptions
- Often difficult practically: Unless code already exists
- Rare with scattered missingness in X and Y
- Important to understand: even if you use another approach

Creating Application-Specific Methods

Creating Application-Specific Methods

- Suppose: $D = \{D_{obs}, D_{mis}\}$ is observed; (M always is)

Creating Application-Specific Methods

- Suppose: $D = \{D_{obs}, D_{mis}\}$ is observed; (M always is)
- Likelihood: $\mathrm{P}(D, M|\theta, \gamma) = \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma)$

Creating Application-Specific Methods

- Suppose: $D = \{D_{obs}, D_{mis}\}$ is observed; (M always is)
- Likelihood: $\mathrm{P}(D, M|\theta, \gamma) = \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma)$
- Quiz: Can we drop $P(M|D, \gamma)$?

Creating Application-Specific Methods

- Suppose: $D = \{D_{obs}, D_{mis}\}$ is observed; (M always is)
- Likelihood: $\mathrm{P}(D, M|\theta, \gamma) = \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma)$
- Quiz: Can we drop $P(M|D, \gamma)$? (MAR: Stoch.& Param. Indep.)

Creating Application-Specific Methods

- Suppose: $D = \{D_{obs}, D_{mis}\}$ is observed; (M always is)
- Likelihood: $\text{P}(D, M|\theta, \gamma) = \text{P}(D|\theta)\text{P}(M|D, \gamma)$
- Quiz: Can we drop $P(M|D, \gamma)$? (MAR: Stoch.& Param. Indep.)
- Suppose now, as usual: D is observed only when M is 1

Creating Application-Specific Methods

- Suppose: $D = \{D_{obs}, D_{mis}\}$ is observed; (M always is)
- Likelihood: $\mathrm{P}(D, M|\theta, \gamma) = \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma)$
- Quiz: Can we drop $P(M|D, \gamma)$? (MAR: Stoch.& Param. Indep.)
- Suppose now, as usual: D is observed only when M is 1
- Likelihood: integrate out the missing observations

Creating Application-Specific Methods

- Suppose: $D = \{D_{obs}, D_{mis}\}$ is observed; (M always is)
- Likelihood: $\mathrm{P}(D, M|\theta, \gamma) = \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma)$
- Quiz: Can we drop $P(M|D, \gamma)$? (MAR: Stoch.& Param. Indep.)
- Suppose now, as usual: D is observed only when M is 1
- Likelihood: integrate out the missing observations

$$P(D_{obs}, M|\theta, \gamma)$$

Creating Application-Specific Methods

- Suppose: $D = \{D_{obs}, D_{mis}\}$ is observed; (M always is)
- Likelihood: $\mathrm{P}(D, M|\theta, \gamma) = \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma)$
- Quiz: Can we drop $P(M|D, \gamma)$? (MAR: Stoch.& Param. Indep.)
- Suppose now, as usual: D is observed only when M is 1
- Likelihood: integrate out the missing observations

$$P(D_{obs}, M|\theta, \gamma) = \int \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma)dD_{mis}$$

Creating Application-Specific Methods

- Suppose: $D = \{D_{obs}, D_{mis}\}$ is observed; (M always is)
- Likelihood: $\mathrm{P}(D, M|\theta, \gamma) = \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma)$
- Quiz: Can we drop $P(M|D, \gamma)$? (MAR: Stoch.& Param. Indep.)
- Suppose now, as usual: D is observed only when M is 1
- Likelihood: integrate out the missing observations

$$\begin{aligned} P(D_{obs}, M|\theta, \gamma) &= \int \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma) dD_{mis} \\ &= P(D_{obs}|\theta)P(M|D_{obs}, \gamma), \end{aligned}$$

Creating Application-Specific Methods

- Suppose: $D = \{D_{obs}, D_{mis}\}$ is observed; (M always is)
- Likelihood: $\mathrm{P}(D, M|\theta, \gamma) = \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma)$
- Quiz: Can we drop $P(M|D, \gamma)$? (MAR: Stoch.& Param. Indep.)
- Suppose now, as usual: D is observed only when M is 1
- Likelihood: integrate out the missing observations

$$\begin{aligned} P(D_{obs}, M|\theta, \gamma) &= \int \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma)dD_{mis} \\ &= P(D_{obs}|\theta)P(M|D_{obs}, \gamma), \\ &\propto P(D_{obs}|\theta) \end{aligned}$$

Creating Application-Specific Methods

- Suppose: $D = \{D_{obs}, D_{mis}\}$ is observed; (M always is)
- Likelihood: $\mathrm{P}(D, M|\theta, \gamma) = \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma)$
- Quiz: Can we drop $P(M|D, \gamma)$? (MAR: Stoch.& Param. Indep.)
- Suppose now, as usual: D is observed only when M is 1
- Likelihood: integrate out the missing observations

$$\begin{aligned} P(D_{obs}, M|\theta, \gamma) &= \int \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma)dD_{mis} \\ &= P(D_{obs}|\theta)P(M|D_{obs}, \gamma), \\ &\propto P(D_{obs}|\theta) \end{aligned}$$

- With MAR: $P(D_{obs}|\theta)$ specialized; no help with X missingness

Creating Application-Specific Methods

- Suppose: $D = \{D_{obs}, D_{mis}\}$ is observed; (M always is)
- Likelihood: $\mathrm{P}(D, M|\theta, \gamma) = \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma)$
- Quiz: Can we drop $P(M|D, \gamma)$? (MAR: Stoch.& Param. Indep.)
- Suppose now, as usual: D is observed only when M is 1
- Likelihood: integrate out the missing observations

$$\begin{aligned} P(D_{obs}, M|\theta, \gamma) &= \int \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma) dD_{mis} \\ &= P(D_{obs}|\theta)P(M|D_{obs}, \gamma), \\ &\propto P(D_{obs}|\theta) \end{aligned}$$

- With MAR: $P(D_{obs}|\theta)$ specialized; no help with X missingness
- Example: Censored Exponential Model (you've seen)

Creating Application-Specific Methods

- Suppose: $D = \{D_{obs}, D_{mis}\}$ is observed; (M always is)
- Likelihood: $\mathrm{P}(D, M|\theta, \gamma) = \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma)$
- Quiz: Can we drop $P(M|D, \gamma)$? (MAR: Stoch.& Param. Indep.)
- Suppose now, as usual: D is observed only when M is 1
- Likelihood: integrate out the missing observations

$$
\begin{aligned}
P(D_{obs}, M|\theta, \gamma) &= \int \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma) dD_{mis} \\
&= P(D_{obs}|\theta)P(M|D_{obs}, \gamma), \\
&\propto P(D_{obs}|\theta)
\end{aligned}
$$

- With MAR: $P(D_{obs}|\theta)$ specialized; no help with X missingness
- Example: Censored Exponential Model (you've seen)
- Without MAR: missingness model is NI, can't be dropped

Creating Application-Specific Methods

- Suppose: $D = \{D_{obs}, D_{mis}\}$ is observed; (M always is)
- Likelihood: $\mathrm{P}(D, M|\theta, \gamma) = \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma)$
- Quiz: Can we drop $P(M|D, \gamma)$? (MAR: Stoch.& Param. Indep.)
- Suppose now, as usual: D is observed only when M is 1
- Likelihood: integrate out the missing observations

$$
\begin{aligned}
P(D_{obs}, M|\theta, \gamma) &= \int \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma) dD_{mis} \\
&= P(D_{obs}|\theta)P(M|D_{obs}, \gamma), \\
&\propto P(D_{obs}|\theta)
\end{aligned}
$$

- With MAR: $P(D_{obs}|\theta)$ specialized; no help with X missingness
- Example: Censored Exponential Model (you've seen)
- Without MAR: missingness model is NI, can't be dropped
- Little known about M: Specifying $P(M|D_{obs}, \gamma)$ hard

Creating Application-Specific Methods

- Suppose: $D = \{D_{obs}, D_{mis}\}$ is observed; (M always is)
- Likelihood: $\mathrm{P}(D, M|\theta, \gamma) = \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma)$
- Quiz: Can we drop $P(M|D, \gamma)$? (MAR: Stoch.& Param. Indep.)
- Suppose now, as usual: D is observed only when M is 1
- Likelihood: integrate out the missing observations

$$\begin{aligned} P(D_{obs}, M|\theta, \gamma) &= \int \mathrm{P}(D|\theta)\mathrm{P}(M|D, \gamma) dD_{mis} \\ &= P(D_{obs}|\theta)P(M|D_{obs}, \gamma), \\ &\propto P(D_{obs}|\theta) \end{aligned}$$

- With MAR: $P(D_{obs}|\theta)$ specialized; no help with X missingness
- Example: Censored Exponential Model (you've seen)
- Without MAR: missingness model is NI, can't be dropped
- Little known about M: Specifying $P(M|D_{obs}, \gamma)$ hard
- NI models: (Heckman, many others) often don't do well when can be evaluated

Multiple Imputation

Multiple Imputation

Point estimates: consistent, efficient; SEs are right (or conservative)

Multiple Imputation

Point estimates: consistent, efficient; SEs are right (or conservative)

- Impute m values for each missing element

Multiple Imputation

Point estimates: consistent, efficient; SEs are right (or conservative)

- Impute m values for each missing element
  - m is bigger with more uncertainty and more missingness

Multiple Imputation

Point estimates: consistent, efficient; SEs are right (or conservative)

- Impute m values for each missing element
    - m is bigger with more uncertainty and more missingness
    - Imputation method (we'll describe later) assumes MAR

Multiple Imputation

Point estimates: consistent, efficient; SEs are right (or conservative)

- Impute m values for each missing element
  - m is bigger with more uncertainty and more missingness
  - Imputation method (we'll describe later) assumes MAR
  - Model must include all (estimation, fundamental) uncertainty

Multiple Imputation

Point estimates: consistent, efficient; SEs are right (or conservative)

- Impute m values for each missing element
  - m is bigger with more uncertainty and more missingness
  - Imputation method (we'll describe later) assumes MAR
  - Model must include all (estimation, fundamental) uncertainty
  - Produces independent imputations

Multiple Imputation

Point estimates: consistent, efficient; SEs are right (or conservative)

- Impute m values for each missing element
  - m is bigger with more uncertainty and more missingness
  - Imputation method (we'll describe later) assumes MAR
  - Model must include all (estimation, fundamental) uncertainty
  - Produces independent imputations
- Create m completed data sets

Multiple Imputation

Point estimates: consistent, efficient; SEs are right (or conservative)

- Impute m values for each missing element
  - m is bigger with more uncertainty and more missingness
  - Imputation method (we'll describe later) assumes MAR
  - Model must include all (estimation, fundamental) uncertainty
  - Produces independent imputations
- Create m completed data sets
  - Observed data are the same across the data sets

Multiple Imputation

Point estimates: consistent, efficient; SEs are right (or conservative)

- Impute m values for each missing element
  - m is bigger with more uncertainty and more missingness
  - Imputation method (we'll describe later) assumes MAR
  - Model must include all (estimation, fundamental) uncertainty
  - Produces independent imputations
- Create m completed data sets
  - Observed data are the same across the data sets
  - Imputations of missing data differ

Multiple Imputation

Point estimates: consistent, efficient; SEs are right (or conservative)

- Impute m values for each missing element
  - m is bigger with more uncertainty and more missingness
  - Imputation method (we'll describe later) assumes MAR
  - Model must include all (estimation, fundamental) uncertainty
  - Produces independent imputations
- Create m completed data sets
  - Observed data are the same across the data sets
  - Imputations of missing data differ
    - Cells we predict well don't differ much

Multiple Imputation

Point estimates: consistent, efficient; SEs are right (or conservative)

- Impute m values for each missing element
  - m is bigger with more uncertainty and more missingness
  - Imputation method (we'll describe later) assumes MAR
  - Model must include all (estimation, fundamental) uncertainty
  - Produces independent imputations
- Create m completed data sets
  - Observed data are the same across the data sets
  - Imputations of missing data differ
    - Cells we predict well don't differ much
    - Cells we can't predict well differ a lot

Multiple Imputation

Point estimates: consistent, efficient; SEs are right (or conservative)

- Impute m values for each missing element
  - m is bigger with more uncertainty and more missingness
  - Imputation method (we'll describe later) assumes MAR
  - Model must include all (estimation, fundamental) uncertainty
  - Produces independent imputations
- Create m completed data sets
  - Observed data are the same across the data sets
  - Imputations of missing data differ
    - Cells we predict well don't differ much
    - Cells we can't predict well differ a lot
- Run whatever statistical method you would have with no missing data for each completed data set

Combining Analyses from Separate Imputed Datasets

Combining Analyses from Separate Imputed Datasets

- By averaging (simple, easy to understand)

Combining Analyses from Separate Imputed Datasets

- By averaging (simple, easy to understand)
  - Point estimate: average individual point estimates, q_j ($j = 1, \ldots, m$)

Combining Analyses from Separate Imputed Datasets

- By averaging (simple, easy to understand)
  - Point estimate: average individual point estimates, q_j ($j = 1, \ldots, m$)

$$\bar{q} = \frac{1}{m} \sum_{j=1}^{m} q_j$$

Combining Analyses from Separate Imputed Datasets

- By averaging (simple, easy to understand)
  - Point estimate: average individual point estimates, q_j ($j = 1, \ldots, m$)

$$\bar{q} = \frac{1}{m} \sum_{j=1}^{m} q_j$$

  - Standard error:

Combining Analyses from Separate Imputed Datasets

- By averaging (simple, easy to understand)
  - Point estimate: average individual point estimates, q_j $(j = 1, \ldots, m)$

$$\bar{q} = \frac{1}{m} \sum_{j=1}^{m} q_j$$

  - Standard error:

$$\mathrm{SE}(q)^2 = \mathrm{mean}(\mathrm{SE}_j^2) + \mathrm{variance}(q_j)\,(1 + 1/m)$$

Combining Analyses from Separate Imputed Datasets

- By averaging (simple, easy to understand)
  - Point estimate: average individual point estimates, q_j ($j = 1, \ldots, m$)

$$\bar{q} = \frac{1}{m} \sum_{j=1}^{m} q_j$$

  - Standard error:

$$\begin{aligned} \mathrm{SE}(q)^2 &= \mathrm{mean}(\mathrm{SE}_j^2) + \mathrm{variance}(q_j)\,(1 + 1/m) \\ &= \text{within} + \text{between} \end{aligned}$$

Combining Analyses from Separate Imputed Datasets

- By averaging (simple, easy to understand)
  - Point estimate: average individual point estimates, q_j ($j = 1, \ldots, m$)

$$\bar{q} = \frac{1}{m} \sum_{j=1}^{m} q_j$$

  - Standard error:

$$\begin{aligned} \mathrm{SE}(q)^2 &= \mathrm{mean}(\mathrm{SE}_j^2) + \mathrm{variance}(q_j)\,(1 + 1/m) \\ &= \text{within} + \text{between} \end{aligned}$$

$(1 + 1/m)$ vanishes as m increases

Combining Analyses from Separate Imputed Datasets

- By averaging (simple, easy to understand)
  - Point estimate: average individual point estimates, q_j $(j = 1, \ldots, m)$

$$\bar{q} = \frac{1}{m} \sum_{j=1}^{m} q_j$$

  - Standard error:

$$\begin{aligned} \mathrm{SE}(q)^2 &= \mathrm{mean}(\mathrm{SE}_j^2) + \mathrm{variance}(q_j)\,(1 + 1/m) \\ &= \mathrm{within} + \mathrm{between} \end{aligned}$$

$(1 + 1/m)$ vanishes as m increases

- By simulation (simpler, fits with our procecures)

Combining Analyses from Separate Imputed Datasets

- By averaging (simple, easy to understand)
  - Point estimate: average individual point estimates, q_j ($j = 1, \ldots, m$)

$$\bar{q} = \frac{1}{m} \sum_{j=1}^{m} q_j$$

  - Standard error:

$$\begin{aligned} \mathrm{SE}(q)^2 &= \mathrm{mean}(\mathrm{SE}_j^2) + \mathrm{variance}(q_j)\,(1 + 1/m) \\ &= \text{within} + \text{between} \end{aligned}$$

$(1 + 1/m)$ vanishes as m increases

- By simulation (simpler, fits with our procecures)
  - draw $1/m$ of needed sims of the QOI from each data set

Combining Analyses from Separate Imputed Datasets

- By averaging (simple, easy to understand)
  - Point estimate: average individual point estimates, q_j ($j = 1, \ldots, m$)

$$\bar{q} = \frac{1}{m} \sum_{j=1}^{m} q_j$$

  - Standard error:

$$\begin{aligned} \mathrm{SE}(q)^2 &= \mathrm{mean}(\mathrm{SE}_j^2) + \mathrm{variance}(q_j)\,(1 + 1/m) \\ &= \text{within} + \text{between} \end{aligned}$$

$(1 + 1/m)$ vanishes as m increases

- By simulation (simpler, fits with our procecures)
  - draw $1/m$ of needed sims of the QOI from each data set
  - combine (i.e., concatenate into a larger set of QOI simulations)

Combining Analyses from Separate Imputed Datasets

- By averaging (simple, easy to understand)
  - Point estimate: average individual point estimates, q_j ($j = 1, \dots, m$)

$$\bar{q} = \frac{1}{m} \sum_{j=1}^{m} q_j$$

  - Standard error:

$$\begin{aligned} \mathrm{SE}(q)^2 &= \mathrm{mean}(\mathrm{SE}_j^2) + \mathrm{variance}(q_j)\,(1 + 1/m) \\ &= \text{within} + \text{between} \end{aligned}$$

$(1 + 1/m)$ vanishes as m increases

- By simulation (simpler, fits with our procecures)
  - draw $1/m$ of needed sims of the QOI from each data set
  - combine (i.e., concatenate into a larger set of QOI simulations)
  - make inferences from combined sims as usual

An Imputation Model

An Imputation Model

- Assume D is complete and: $\underset{N\times 1}{D_i} \sim N(D_i|\mu,\Sigma)$ (SURM w/o X)

An Imputation Model

- Assume D is complete and: $\underset{N\times 1}{D_i} \sim N(D_i|\mu, \Sigma) \quad$ (SURM w/o X)
- Seems crazy, but will imply regression of each var on all others

An Imputation Model

- Assume D is complete and: $\underset{N\times 1}{D_i} \sim N(D_i|\mu,\Sigma)$ (SURM w/o X)
- Seems crazy, but will imply regression of each var on all others
- Likelihood: $L(\mu,\Sigma|D) \propto \prod_{i=1}^{n} N(D_i|\mu,\Sigma)$

An Imputation Model

- Assume D is complete and: $\underset{N\times 1}{D_i} \sim N(D_i|\mu,\Sigma)$ (SURM w/o X)
- Seems crazy, but will imply regression of each var on all others
- Likelihood: $L(\mu,\Sigma|D) \propto \prod_{i=1}^{n} N(D_i|\mu,\Sigma)$
- With D partially missing

An Imputation Model

- Assume D is complete and: $\underset{N\times 1}{D_i} \sim N(D_i|\mu,\Sigma)$ (SURM w/o X)
- Seems crazy, but will imply regression of each var on all others
- Likelihood: $L(\mu,\Sigma|D) \propto \prod_{i=1}^{n} N(D_i|\mu,\Sigma)$
- With D partially missing

$$L(\mu,\Sigma|D_{obs})$$

An Imputation Model

- Assume D is complete and: $\underset{N\times 1}{D_i} \sim N(D_i|\mu, \Sigma)$ (SURM w/o X)
- Seems crazy, but will imply regression of each var on all others
- Likelihood: $L(\mu, \Sigma|D) \propto \prod_{i=1}^{n} N(D_i|\mu, \Sigma)$
- With D partially missing

$$L(\mu, \Sigma|D_{obs}) \propto \prod_{i=1}^{n} \int N(D_i|\mu, \Sigma) dD_{mis}$$

An Imputation Model

- Assume D is complete and: $\underset{N\times 1}{D_i} \sim N(D_i|\mu,\Sigma)$ (SURM w/o X)
- Seems crazy, but will imply regression of each var on all others
- Likelihood: $L(\mu,\Sigma|D) \propto \prod_{i=1}^{n} N(D_i|\mu,\Sigma)$
- With D partially missing

$$
\begin{aligned}
L(\mu,\Sigma|D_{obs}) &\propto \prod_{i=1}^{n} \int N(D_i|\mu,\Sigma)\, dD_{mis} \\
&= \prod_{i=1}^{n} N(D_{i,obs}|\mu_{obs},\Sigma_{obs})
\end{aligned}
$$

An Imputation Model

- Assume D is complete and: $\underset{N\times 1}{D_i} \sim N(D_i|\mu,\Sigma)$ (SURM w/o X)
- Seems crazy, but will imply regression of each var on all others
- Likelihood: $L(\mu,\Sigma|D) \propto \prod_{i=1}^{n} N(D_i|\mu,\Sigma)$
- With D partially missing

$$
\begin{aligned}
L(\mu,\Sigma|D_{obs}) &\propto \prod_{i=1}^{n} \int N(D_i|\mu,\Sigma)dD_{mis} \\
&= \prod_{i=1}^{n} N(D_{i,obs}|\mu_{obs},\Sigma_{obs})
\end{aligned}
$$

since marginals of MVN's are normal.

An Imputation Model

- Assume D is complete and: $\underset{N\times 1}{D_i} \sim N(D_i|\mu, \Sigma)$ (SURM w/o X)
- Seems crazy, but will imply regression of each var on all others
- Likelihood: $L(\mu, \Sigma|D) \propto \prod_{i=1}^{n} N(D_i|\mu, \Sigma)$
- With D partially missing

$$
\begin{aligned}
L(\mu, \Sigma|D_{obs}) &\propto \prod_{i=1}^{n} \int N(D_i|\mu, \Sigma)\, dD_{mis} \\
&= \prod_{i=1}^{n} N(D_{i,obs}|\mu_{obs}, \Sigma_{obs})
\end{aligned}
$$

  since marginals of MVN's are normal.
- Simple theoretically: merely a likelihood model for data (D_{obs}, M) and same parameters as when fully observed (μ, Σ)

An Imputation Model

- Assume D is complete and: $\underset{N\times 1}{D_i} \sim N(D_i|\mu,\Sigma)$ (SURM w/o X)
- Seems crazy, but will imply regression of each var on all others
- Likelihood: $L(\mu,\Sigma|D) \propto \prod_{i=1}^{n} N(D_i|\mu,\Sigma)$
- With D partially missing

$$\begin{aligned} L(\mu,\Sigma|D_{obs}) &\propto \prod_{i=1}^{n} \int N(D_i|\mu,\Sigma)dD_{mis} \\ &= \prod_{i=1}^{n} N(D_{i,obs}|\mu_{obs},\Sigma_{obs}) \end{aligned}$$

  since marginals of MVN's are normal.
- Simple theoretically: merely a likelihood model for data (D_{obs}, M) and same parameters as when fully observed (μ,Σ)
- Complicated computationally: $D_{i,obs}$ has different elements observed for each i; each observation is informative about different pieces of (μ,Σ)

An Imputation Model (continued)

- Hard Statistically: params increase fast in number of vars (p)

An Imputation Model (continued)

- Hard Statistically: params increase fast in number of vars (p)

$$\text{parameters} = \text{parameters}(\mu) + \text{parameters}(\Sigma)$$

An Imputation Model (continued)

- Hard Statistically: params increase fast in number of vars (p)

$$\begin{aligned} \text{parameters} &= \text{parameters}(\mu) + \text{parameters}(\Sigma) \\ &= p + p(p+1)/2 \end{aligned}$$

An Imputation Model (continued)

- Hard Statistically: params increase fast in number of vars (p)

$$\begin{aligned} \text{parameters} &= \text{parameters}(\mu) + \text{parameters}(\Sigma) \\ &= p + p(p+1)/2 = p(p+3)/2. \end{aligned}$$

An Imputation Model (continued)

- Hard Statistically: params increase fast in number of vars (p)

$$\begin{aligned} \text{parameters} &= \text{parameters}(\mu) + \text{parameters}(\Sigma) \\ &= p + p(p+1)/2 = p(p+3)/2. \end{aligned}$$

E.g., for $p = 5$, parameters$= 20$; for $p = 40$ parameters$= 860$ (with what sample size?)

An Imputation Model (continued)

- Hard Statistically: params increase fast in number of vars (p)

$$\begin{aligned}\text{parameters} &= \text{parameters}(\mu) + \text{parameters}(\Sigma)\\ &= p + p(p+1)/2 = p(p+3)/2.\end{aligned}$$

  E.g., for $p = 5$, parameters= 20; for $p = 40$ parameters= 860 (with what sample size?)

- Specialized Models, such as for categorical or mixed variables, are harder to apply and do not usually perform better

An Imputation Model (continued)

- Hard Statistically: params increase fast in number of vars (p)

$$\begin{aligned} \text{parameters} &= \text{parameters}(\mu) + \text{parameters}(\Sigma) \\ &= p + p(p+1)/2 = p(p+3)/2. \end{aligned}$$

  E.g., for $p = 5$, parameters$= 20$; for $p = 40$ parameters$= 860$ (with what sample size?)
- Specialized Models, such as for categorical or mixed variables, are harder to apply and do not usually perform better
- For social science survey data (mostly ordinal scales): this is a reasonable choice for imputation, even if not for analysis

Creating Imputations From This Model

Creating Imputations From This Model

- Suppose $D = \{Y, X\}$ with 2 vars; missingness only in Y

Creating Imputations From This Model

- Suppose $D = \{Y, X\}$ with 2 vars; missingness only in Y
- $\rightsquigarrow D$ is bivariate normal

Creating Imputations From This Model

- Suppose $D = \{Y, X\}$ with 2 vars; missingness only in Y
- $\rightsquigarrow D$ is bivariate normal

$$D \sim N(D|\mu, \Sigma)$$

Creating Imputations From This Model

- Suppose $D = \{Y, X\}$ with 2 vars; missingness only in Y
- $\rightsquigarrow$ D is bivariate normal

$$D \sim N(D|\mu, \Sigma) = N\left[\begin{pmatrix} Y \\ X \end{pmatrix} \middle| \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix}\right]$$

Creating Imputations From This Model

- Suppose $D = \{Y, X\}$ with 2 vars; missingness only in Y
- $\rightsquigarrow D$ is bivariate normal

$$D \sim N(D|\mu, \Sigma) = N\left[\begin{pmatrix} Y \\ X \end{pmatrix} \middle| \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix}\right]$$

- Conditionals of bivariate normals are normal

$$Y|X \sim N\left[y|E(Y|X), V(Y|X)\right]$$

Creating Imputations From This Model

- Suppose $D = \{Y, X\}$ with 2 vars; missingness only in Y
- $\rightsquigarrow$ D is bivariate normal

$$D \sim N(D|\mu, \Sigma) = N\left[\begin{pmatrix} Y \\ X \end{pmatrix} \middle| \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix}\right]$$

- Conditionals of bivariate normals are normal

$$Y|X \sim N\left[y|E(Y|X), V(Y|X)\right]$$

- CEF is regression of Y on all X's (recall: SURM has no X)

$$E(Y|X) = \mu_y + \beta(X - \mu_x)$$

Creating Imputations From This Model

- Suppose $D = \{Y, X\}$ with 2 vars; missingness only in Y
- $\rightsquigarrow$ D is bivariate normal

$$D \sim N(D|\mu, \Sigma) = N\left[\begin{pmatrix} Y \\ X \end{pmatrix} \middle| \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix}\right]$$

- Conditionals of bivariate normals are normal

$$Y|X \sim N\left[y|E(Y|X), V(Y|X)\right]$$

- CEF is regression of Y on all X's (recall: SURM has no X)

$$E(Y|X) = \mu_y + \beta(X - \mu_x)$$

- Other details: $\beta = \sigma_{xy}/\sigma_x, \quad V(Y|X) = \sigma_y - \sigma^2_{xy}/\sigma_x$

Creating Imputations From This Model

- Suppose $D = \{Y, X\}$ with 2 vars; missingness only in Y
- $\rightsquigarrow$ D is bivariate normal

$$D \sim N(D|\mu, \Sigma) = N\left[\begin{pmatrix} Y \\ X \end{pmatrix} \middle| \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix}\right]$$

- Conditionals of bivariate normals are normal

$$Y|X \sim N\left[y|E(Y|X), V(Y|X)\right]$$

- CEF is regression of Y on all X's (recall: SURM has no X)

$$E(Y|X) = \mu_y + \beta(X - \mu_x)$$

- Other details: $\beta = \sigma_{xy}/\sigma_x, \quad V(Y|X) = \sigma_y - \sigma_{xy}^2/\sigma_x$
- To simulate:

Creating Imputations From This Model

- Suppose $D = \{Y, X\}$ with 2 vars; missingness only in Y
- $\rightsquigarrow$ D is bivariate normal

$$D \sim N(D|\mu, \Sigma) = N\left[\begin{pmatrix} Y \\ X \end{pmatrix} \middle| \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix}\right]$$

- Conditionals of bivariate normals are normal

$$Y|X \sim N\left[y|E(Y|X), V(Y|X)\right]$$

- CEF is regression of Y on all X's (recall: SURM has no X)

$$E(Y|X) = \mu_y + \beta(X - \mu_x)$$

- Other details: $\beta = \sigma_{xy}/\sigma_x, \quad V(Y|X) = \sigma_y - \sigma_{xy}^2/\sigma_x$
- To simulate:
  - Maximize likelihood; draw sims of $\tilde{\mu}$ and $\tilde{\Sigma}$

Creating Imputations From This Model

- Suppose $D = \{Y, X\}$ with 2 vars; missingness only in Y
- $\rightsquigarrow$ D is bivariate normal

$$D \sim N(D|\mu, \Sigma) = N\left[\begin{pmatrix} Y \\ X \end{pmatrix} \middle| \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix}\right]$$

- Conditionals of bivariate normals are normal

$$Y|X \sim N\left[y|E(Y|X), V(Y|X)\right]$$

- CEF is regression of Y on all X's (recall: SURM has no X)

$$E(Y|X) = \mu_y + \beta(X - \mu_x)$$

- Other details: $\beta = \sigma_{xy}/\sigma_x, \quad V(Y|X) = \sigma_y - \sigma_{xy}^2/\sigma_x$
- To simulate:
  - Maximize likelihood; draw sims of $\tilde{\mu}$ and $\tilde{\Sigma}$
  - Calculate $\tilde{\beta}$, $\tilde{\sigma}$s; then sims of $E(Y|X)$ and $V(Y|X)$

Creating Imputations From This Model

- Suppose $D = \{Y, X\}$ with 2 vars; missingness only in Y
- $\rightsquigarrow$ D is bivariate normal

$$D \sim N(D|\mu, \Sigma) = N\left[\begin{pmatrix} Y \\ X \end{pmatrix} \middle| \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix}\right]$$

- Conditionals of bivariate normals are normal

$$Y|X \sim N\left[y|E(Y|X), V(Y|X)\right]$$

- CEF is regression of Y on all X's (recall: SURM has no X)

$$E(Y|X) = \mu_y + \beta(X - \mu_x)$$

- Other details: $\beta = \sigma_{xy}/\sigma_x, \quad V(Y|X) = \sigma_y - \sigma^2_{xy}/\sigma_x$
- To simulate:
  - Maximize likelihood; draw sims of $\tilde{\mu}$ and $\tilde{\Sigma}$
  - Calculate $\tilde{\beta}$, $\tilde{\sigma}$s; then sims of $E(Y|X)$ and $V(Y|X)$
  - Draw a simulation of the missing Y from the conditional normal

Creating Imputations From This Model

- Suppose $D = \{Y, X\}$ with 2 vars; missingness only in Y
- $\rightsquigarrow D$ is bivariate normal

$$D \sim N(D|\mu, \Sigma) = N\left[\begin{pmatrix} Y \\ X \end{pmatrix} \middle| \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix}\right]$$

- Conditionals of bivariate normals are normal

$$Y|X \sim N\left[y|E(Y|X), V(Y|X)\right]$$

- CEF is regression of Y on all X's (recall: SURM has no X)

$$E(Y|X) = \mu_y + \beta(X - \mu_x)$$

- Other details: $\beta = \sigma_{xy}/\sigma_x, \quad V(Y|X) = \sigma_y - \sigma_{xy}^2/\sigma_x$
- To simulate:
  - Maximize likelihood; draw sims of $\tilde{\mu}$ and $\tilde{\Sigma}$
  - Calculate $\tilde{\beta}$, $\tilde{\sigma}$s; then sims of $E(Y|X)$ and $V(Y|X)$
  - Draw a simulation of the missing Y from the conditional normal
  - Equivalent to drawing from $\tilde{y}_i = X\tilde{\beta} + \tilde{\epsilon}_i$ (estimation, fundamental uncertainty)

Overview

Missingness Assumptions

Application Specific Methods

Multiple Imputation

Computational Algorithms

What Can Go Wrong

Time Series, Cross-Sectional Imputations

Computational Algorithms: Optim, EM, EMs (p.1 of 3)

Computational Algorithms: Optim, EM, EMs (p.1 of 3)

- Optim with hundreds of parameters would work very slowly

Computational Algorithms: Optim, EM, EMs (p.1 of 3)

- Optim with hundreds of parameters would work very slowly
- EM (expectation maximization): fast algorithm for maximum

Computational Algorithms: Optim, EM, EMs (p.1 of 3)

- Optim with hundreds of parameters would work very slowly
- EM (expectation maximization): fast algorithm for maximum
    - Much faster than optim

Computational Algorithms: Optim, EM, EMs (p.1 of 3)

- Optim with hundreds of parameters would work very slowly
- EM (expectation maximization): fast algorithm for maximum
  - Much faster than optim
  - Intuition:

Computational Algorithms: Optim, EM, EMs (p.1 of 3)

- Optim with hundreds of parameters would work very slowly
- EM (expectation maximization): fast algorithm for maximum
  - Much faster than optim
  - Intuition:
    - Without missingness, estimating β is easy: run LS

Computational Algorithms: Optim, EM, EMs (p.1 of 3)

- Optim with hundreds of parameters would work very slowly
- EM (expectation maximization): fast algorithm for maximum
  - Much faster than optim
  - Intuition:
    - Without missingness, estimating β is easy: run LS
    - If β is known, imputation is easy: draw $\tilde{\epsilon}$, use $\tilde{y} = x\beta + \tilde{\epsilon}$

Computational Algorithms: Optim, EM, EMs (p.1 of 3)

- Optim with hundreds of parameters would work very slowly
- EM (expectation maximization): fast algorithm for maximum
  - Much faster than optim
  - Intuition:
    - Without missingness, estimating β is easy: run LS
    - If β is known, imputation is easy: draw $\tilde{\epsilon}$, use $\tilde{y} = x\beta + \tilde{\epsilon}$
  - EM works by iterating between

Computational Algorithms: Optim, EM, EMs (p.1 of 3)

- Optim with hundreds of parameters would work very slowly
- EM (expectation maximization): fast algorithm for maximum
  - Much faster than optim
  - Intuition:
    - Without missingness, estimating β is easy: run LS
    - If β is known, imputation is easy: draw $\tilde{\epsilon}$, use $\tilde{y} = x\beta + \tilde{\epsilon}$
  - EM works by iterating between
    - Impute $\hat{Y}$ with $x\hat{\beta}$, given current estimates, $\hat{\beta}$

Computational Algorithms: Optim, EM, EMs (p.1 of 3)

- Optim with hundreds of parameters would work very slowly
- EM (expectation maximization): fast algorithm for maximum
  - Much faster than optim
  - Intuition:
    - Without missingness, estimating β is easy: run LS
    - If β is known, imputation is easy: draw $\tilde{\epsilon}$, use $\tilde{y} = x\beta + \tilde{\epsilon}$
  - EM works by iterating between
    - Impute $\hat{Y}$ with $x\hat{\beta}$, given current estimates, $\hat{\beta}$
    - Estimate β (by LS) with current imputations of Y

Computational Algorithms: Optim, EM, EMs (p.1 of 3)

- Optim with hundreds of parameters would work very slowly
- EM (expectation maximization): fast algorithm for maximum
  - Much faster than optim
  - Intuition:
    - Without missingness, estimating β is easy: run LS
    - If β is known, imputation is easy: draw $\tilde{\epsilon}$, use $\tilde{y} = x\beta + \tilde{\epsilon}$
  - EM works by iterating between
    - Impute $\hat{Y}$ with $x\hat{\beta}$, given current estimates, $\hat{\beta}$
    - Estimate β (by LS) with current imputations of Y
  - Can easily do imputation via $x\hat{\beta} + \tilde{\epsilon}$

Computational Algorithms: Optim, EM, EMs (p.1 of 3)

- Optim with hundreds of parameters would work very slowly
- EM (expectation maximization): fast algorithm for maximum
  - Much faster than optim
  - Intuition:
    - Without missingness, estimating β is easy: run LS
    - If β is known, imputation is easy: draw $\tilde{\epsilon}$, use $\tilde{y} = x\beta + \tilde{\epsilon}$
  - EM works by iterating between
    - Impute $\hat{Y}$ with $x\hat{\beta}$, given current estimates, $\hat{\beta}$
    - Estimate β (by LS) with current imputations of Y
  - Can easily do imputation via $x\hat{\beta} + \tilde{\epsilon}$
  - Problem: SEs too small due to no estimation uncertainty ($\hat{\beta} \neq \beta$); i.e., we need to draw β from its posterior first

Computational Algorithms: Optim, EM, EMs (p.1 of 3)

- Optim with hundreds of parameters would work very slowly
- EM (expectation maximization): fast algorithm for maximum
  - Much faster than optim
  - Intuition:
    - Without missingness, estimating β is easy: run LS
    - If β is known, imputation is easy: draw $\tilde{\epsilon}$, use $\tilde{y} = x\beta + \tilde{\epsilon}$
  - EM works by iterating between
    - Impute $\hat{Y}$ with $x\hat{\beta}$, given current estimates, $\hat{\beta}$
    - Estimate β (by LS) with current imputations of Y
  - Can easily do imputation via $x\hat{\beta} + \tilde{\epsilon}$
  - Problem: SEs too small due to no estimation uncertainty ($\hat{\beta} \neq \beta$); i.e., we need to draw β from its posterior first
- EMs: EM for param maximization and then simulation

Computational Algorithms: Optim, EM, EMs (p.1 of 3)

- Optim with hundreds of parameters would work very slowly
- EM (expectation maximization): fast algorithm for maximum
  - Much faster than optim
  - Intuition:
    - Without missingness, estimating β is easy: run LS
    - If β is known, imputation is easy: draw $\tilde{\epsilon}$, use $\tilde{y} = x\beta + \tilde{\epsilon}$
  - EM works by iterating between
    - Impute $\hat{Y}$ with $x\hat{\beta}$, given current estimates, $\hat{\beta}$
    - Estimate β (by LS) with current imputations of Y
  - Can easily do imputation via $x\hat{\beta} + \tilde{\epsilon}$
  - Problem: SEs too small due to no estimation uncertainty ($\hat{\beta} \neq \beta$); i.e., we need to draw β from its posterior first
- EMs: EM for param maximization and then simulation
  - Add estimation uncertainty by drawing $\tilde{\beta} \sim N(\hat{\beta}, \hat{V}(\hat{\beta}))$

Computational Algorithms: Optim, EM, EMs (p.1 of 3)

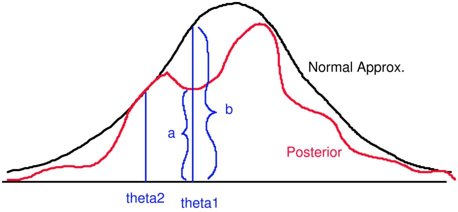
- Optim with hundreds of parameters would work very slowly
- EM (expectation maximization): fast algorithm for maximum
  - Much faster than optim
  - Intuition:
    - Without missingness, estimating β is easy: run LS
    - If β is known, imputation is easy: draw $\tilde{\epsilon}$, use $\tilde{y} = x\beta + \tilde{\epsilon}$
  - EM works by iterating between
    - Impute $\hat{Y}$ with $x\hat{\beta}$, given current estimates, $\hat{\beta}$
    - Estimate β (by LS) with current imputations of Y
  - Can easily do imputation via $x\hat{\beta} + \tilde{\epsilon}$
  - Problem: SEs too small due to no estimation uncertainty ($\hat{\beta} \neq \beta$); i.e., we need to draw β from its posterior first
- EMs: EM for param maximization and then simulation
  - Add estimation uncertainty by drawing $\tilde{\beta} \sim N(\hat{\beta}, \hat{V}(\hat{\beta}))$
  - The central limit theorem guarantees that this works as $n \longrightarrow \infty$, but for real sample sizes it may be inadequate

Computational Algorithms: EMis, EMB (p.2 of 3)

- EMis: EM with simulation via importance resampling (probabilistic rejection sampling to draw from the posterior)

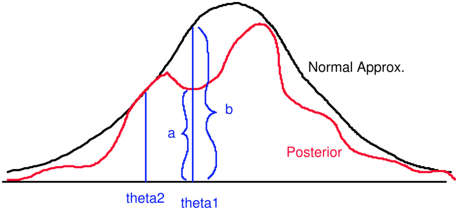
Computational Algorithms: EMis, EMB (p.2 of 3)

- EMis: EM with simulation via importance resampling (probabilistic rejection sampling to draw from the posterior)



Computational Algorithms: EMis, EMB (p.2 of 3)

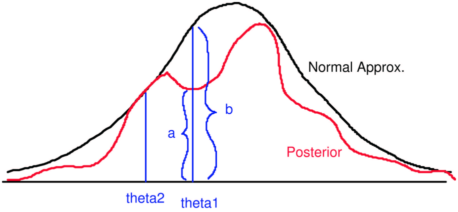
- EMis: EM with simulation via importance resampling (probabilistic rejection sampling to draw from the posterior)



Keep $\tilde{\theta}_1$ with probability $\propto a/b$ (the importance ratio).
Keep $\tilde{\theta}_2$ with probability 1.

Computational Algorithms: EMis, EMB (p.2 of 3)

- EMis: EM with simulation via importance resampling (probabilistic rejection sampling to draw from the posterior)

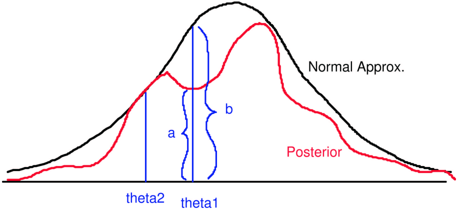


Keep $\tilde{\theta}_1$ with probability $\propto a/b$ (the importance ratio).
Keep $\tilde{\theta}_2$ with probability 1.

- EMB: EM With Bootstrap

Computational Algorithms: EMis, EMB (p.2 of 3)

- EMis: EM with simulation via importance resampling (probabilistic rejection sampling to draw from the posterior)

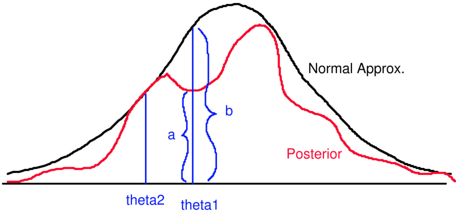


Keep $\tilde{\theta}_1$ with probability $\propto a/b$ (the importance ratio).
Keep $\tilde{\theta}_2$ with probability 1.

- EMB: EM With Bootstrap
  - Draw m sets of n obs (with replacement) from the data

Computational Algorithms: EMis, EMB (p.2 of 3)

- EMis: EM with simulation via importance resampling (probabilistic rejection sampling to draw from the posterior)

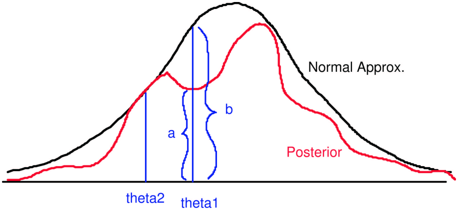


Keep $\tilde{\theta}_1$ with probability $\propto a/b$ (the importance ratio).
Keep $\tilde{\theta}_2$ with probability 1.

- EMB: EM With Bootstrap
  - Draw m sets of n obs (with replacement) from the data
  - Use EM to estimate β and Σ in each (for estimation uncertainty)

Computational Algorithms: EMis, EMB (p.2 of 3)

- EMis: EM with simulation via importance resampling (probabilistic rejection sampling to draw from the posterior)

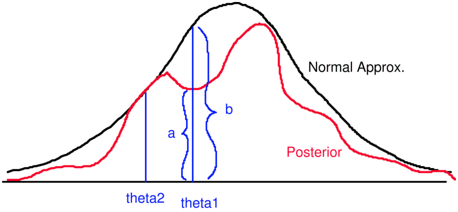


Keep $\tilde{\theta}_1$ with probability $\propto a/b$ (the importance ratio).
Keep $\tilde{\theta}_2$ with probability 1.

- EMB: EM With Bootstrap
  - Draw m sets of n obs (with replacement) from the data
  - Use EM to estimate β and Σ in each (for estimation uncertainty)
  - Impute D_{mis} from the model (for fundamental uncertainty)

Computational Algorithms: EMis, EMB (p.2 of 3)

- EMis: EM with simulation via importance resampling (probabilistic rejection sampling to draw from the posterior)

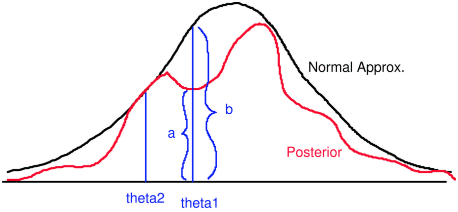


Keep $\tilde{\theta}_1$ with probability $\propto a/b$ (the importance ratio).
Keep $\tilde{\theta}_2$ with probability 1.

- EMB: EM With Bootstrap
  - Draw m sets of n obs (with replacement) from the data
  - Use EM to estimate β and Σ in each (for estimation uncertainty)
  - Impute D_{mis} from the model (for fundamental uncertainty)
  - Lightning fast; works with very large data sets

Computational Algorithms: EMis, EMB (p.2 of 3)

- EMis: EM with simulation via importance resampling (probabilistic rejection sampling to draw from the posterior)



Keep $\tilde{\theta}_1$ with probability $\propto a/b$ (the importance ratio).
Keep $\tilde{\theta}_2$ with probability 1.

- EMB: EM With Bootstrap
  - Draw m sets of n obs (with replacement) from the data
  - Use EM to estimate β and Σ in each (for estimation uncertainty)
  - Impute D_{mis} from the model (for fundamental uncertainty)
  - Lightning fast; works with very large data sets
  - Basis for Amelia II

Computational Algorithms: MCMC, IP (p.3 of 3)

Computational Algorithms: MCMC, IP (p.3 of 3)

- The IP algorithm: Repeat until convergence (≈ stochastic EM)

Computational Algorithms: MCMC, IP (p.3 of 3)

- The IP algorithm: Repeat until convergence (≈ stochastic EM)
  - I-Step: draw D_{mis} from $P(D_{mis}|D_{obs}, \tilde{\theta})$ (i.e., $\tilde{y} = x\tilde{\beta} + \tilde{\epsilon}$)

Computational Algorithms: MCMC, IP (p.3 of 3)

- The IP algorithm: Repeat until convergence (≈ stochastic EM)
  - I-Step: draw D_{mis} from $P(D_{mis}|D_{obs},\tilde{\theta})$ (i.e., $\tilde{y} = x\tilde{\beta} + \tilde{\epsilon}$)
  - P-Step: draw θ from $P(\theta|D_{obs},\tilde{D}_{mis})$

Computational Algorithms: MCMC, IP (p.3 of 3)

- The IP algorithm: Repeat until convergence ($\approx$ stochastic EM)
  - I-Step: draw D_{mis} from $P(D_{mis}|D_{obs}, \tilde{\theta})$ (i.e., $\tilde{y} = x\tilde{\beta} + \tilde{\epsilon}$)
  - P-Step: draw θ from $P(\theta|D_{obs}, \tilde{D}_{mis})$
- IP is an example of MCMC (Markov Chain Monte Carlo)

Computational Algorithms: MCMC, IP (p.3 of 3)

- The IP algorithm: Repeat until convergence (≈ stochastic EM)
  - I-Step: draw D_{mis} from $P(D_{mis}|D_{obs}, \tilde{\theta})$ (i.e., $\tilde{y} = x\tilde{\beta} + \tilde{\epsilon}$)
  - P-Step: draw θ from $P(\theta|D_{obs}, \tilde{D}_{mis})$
- IP is an example of MCMC (Markov Chain Monte Carlo)
  - 1 of 1990+'s most important developments in stats

Computational Algorithms: MCMC, IP (p.3 of 3)

- The IP algorithm: Repeat until convergence ($\approx$ stochastic EM)
  - I-Step: draw D_{mis} from $P(D_{mis}|D_{obs}, \tilde{\theta})$ (i.e., $\tilde{y} = x\tilde{\beta} + \tilde{\epsilon}$)
  - P-Step: draw θ from $P(\theta|D_{obs}, \tilde{D}_{mis})$
- IP is an example of MCMC (Markov Chain Monte Carlo)
  - 1 of 1990+'s most important developments in stats
  - enabled statisticians to do things they never dreamed possible

Computational Algorithms: MCMC, IP (p.3 of 3)

- The IP algorithm: Repeat until convergence (≈ stochastic EM)
  - I-Step: draw D_{mis} from $P(D_{mis}|D_{obs}, \tilde{\theta})$ (i.e., $\tilde{y} = x\tilde{\beta} + \tilde{\epsilon}$)
  - P-Step: draw θ from $P(\theta|D_{obs}, \tilde{D}_{mis})$
- IP is an example of MCMC (Markov Chain Monte Carlo)
  - 1 of 1990+'s most important developments in stats
  - enabled statisticians to do things they never dreamed possible
  - requires considerable expertise; doesn't help others as much

Computational Algorithms: MCMC, IP (p.3 of 3)

- The IP algorithm: Repeat until convergence ($\approx$ stochastic EM)
  - I-Step: draw D_{mis} from $P(D_{mis}|D_{obs}, \tilde{\theta})$ (i.e., $\tilde{y} = x\tilde{\beta} + \tilde{\epsilon}$)
  - P-Step: draw θ from $P(\theta|D_{obs}, \tilde{D}_{mis})$
- IP is an example of MCMC (Markov Chain Monte Carlo)
  - 1 of 1990+'s most important developments in stats
  - enabled statisticians to do things they never dreamed possible
  - requires considerable expertise; doesn't help others as much
  - Few MCMC routines are in canned packages

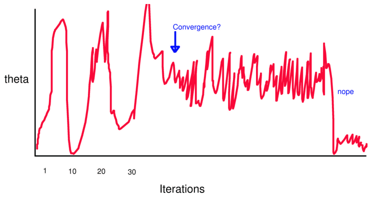
Computational Algorithms: MCMC, IP (p.3 of 3)

- The IP algorithm: Repeat until convergence (≈ stochastic EM)
  - I-Step: draw D_{mis} from $P(D_{mis}|D_{obs}, \tilde{\theta})$ (i.e., $\tilde{y} = x\tilde{\beta} + \tilde{\epsilon}$)
  - P-Step: draw θ from $P(\theta|D_{obs}, \tilde{D}_{mis})$
- IP is an example of MCMC (Markov Chain Monte Carlo)
  - 1 of 1990+'s most important developments in stats
  - enabled statisticians to do things they never dreamed possible
  - requires considerable expertise; doesn't help others as much
  - Few MCMC routines are in canned packages
- Hard to know when MCMC algorithms converge

Computational Algorithms: MCMC, IP (p.3 of 3)

- The IP algorithm: Repeat until convergence (≈ stochastic EM)
  - I-Step: draw D_{mis} from $P(D_{mis}|D_{obs}, \tilde{\theta})$ (i.e., $\tilde{y} = x\tilde{\beta} + \tilde{\epsilon}$)
  - P-Step: draw θ from $P(\theta|D_{obs}, \tilde{D}_{mis})$
- IP is an example of MCMC (Markov Chain Monte Carlo)
  - 1 of 1990+'s most important developments in stats
  - enabled statisticians to do things they never dreamed possible
  - requires considerable expertise; doesn't help others as much
  - Few MCMC routines are in canned packages
- Hard to know when MCMC algorithms converge

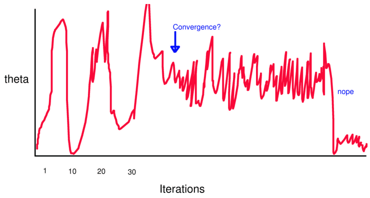
- Convergence is asymptotic



Computational Algorithms: MCMC, IP (p.3 of 3)

- The IP algorithm: Repeat until convergence ($\approx$ stochastic EM)
  - I-Step: draw D_{mis} from $P(D_{mis}|D_{obs}, \tilde{\theta})$ (i.e., $\tilde{y} = x\tilde{\beta} + \tilde{\epsilon}$)
  - P-Step: draw θ from $P(\theta|D_{obs}, \tilde{D}_{mis})$
- IP is an example of MCMC (Markov Chain Monte Carlo)
  - 1 of 1990+'s most important developments in stats
  - enabled statisticians to do things they never dreamed possible
  - requires considerable expertise; doesn't help others as much
  - Few MCMC routines are in canned packages
- Hard to know when MCMC algorithms converge

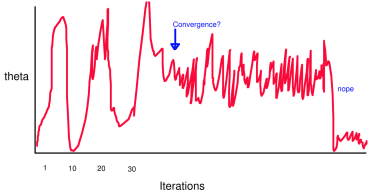
- Convergence is asymptotic
- Plot traces hard to interpret



Computational Algorithms: MCMC, IP (p.3 of 3)

- The IP algorithm: Repeat until convergence ($\approx$ stochastic EM)
  - I-Step: draw D_{mis} from $P(D_{mis}|D_{obs}, \tilde{\theta})$ (i.e., $\tilde{y} = x\tilde{\beta} + \tilde{\epsilon}$)
  - P-Step: draw θ from $P(\theta|D_{obs}, \tilde{D}_{mis})$
- IP is an example of MCMC (Markov Chain Monte Carlo)
  - 1 of 1990+'s most important developments in stats
  - enabled statisticians to do things they never dreamed possible
  - requires considerable expertise; doesn't help others as much
  - Few MCMC routines are in canned packages
- Hard to know when MCMC algorithms converge

- Convergence is asymptotic
- Plot traces hard to interpret
- Worst-converging parameter controls the system

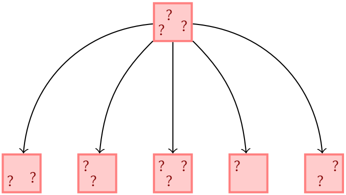


Multiple Imputation: Amelia Style

Multiple Imputation: Amelia Style

incomplete data

Multiple Imputation: Amelia Style

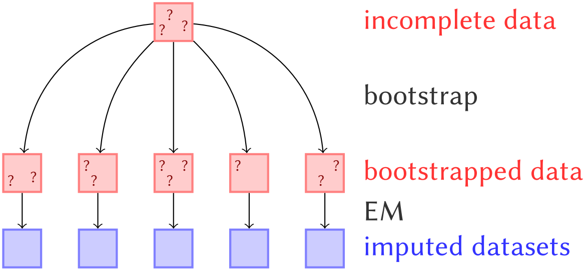


incomplete data

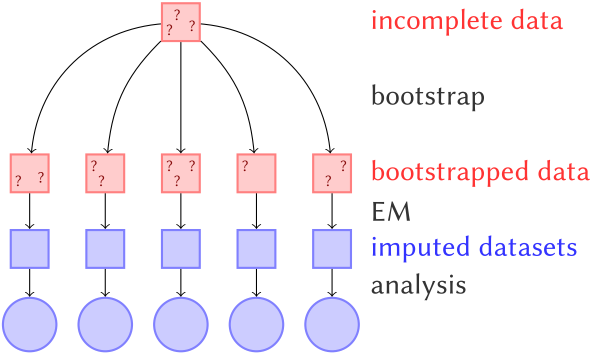
bootstrap

bootstrapped data

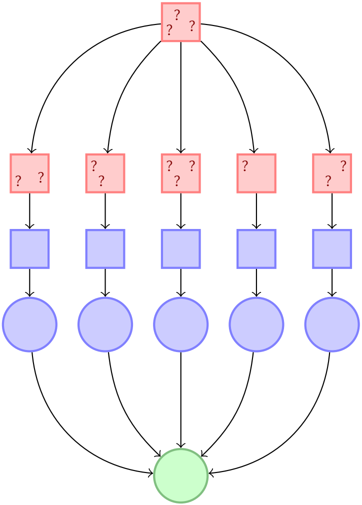
Multiple Imputation: Amelia Style



Multiple Imputation: Amelia Style



Multiple Imputation: Amelia Style



incomplete data

bootstrap

bootstrapped data

EM

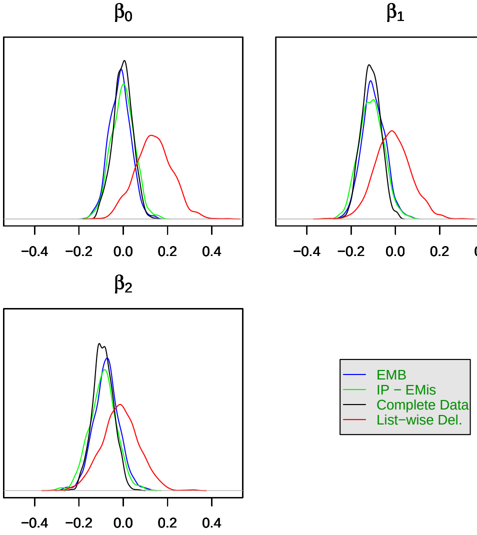
imputed datasets

analysis

combination

final results

Comparisons of Posterior Density Approximations



EMB: similar results, much faster, far easier to program

What Can Go Wrong and What to Do

What Can Go Wrong and What to Do

- Risk: Inference is learning about facts we don't have with facts we have; we assume the 2 are related!

What Can Go Wrong and What to Do

- Risk: Inference is learning about facts we don't have with facts we have; we assume the 2 are related!
- Robustness: Imputation and analysis separated $\rightsquigarrow$ imputation affects only missing observations; High missingness reduces the property

What Can Go Wrong and What to Do

- Risk: Inference is learning about facts we don't have with facts we have; we assume the 2 are related!
- Robustness: Imputation and analysis separated $\rightsquigarrow$ imputation affects only missing observations; High missingness reduces the property
- Vars to include: At least as much information in the imputation model as in the analysis model: all vars in analysis model; others that would help predict (e.g., All measures of a variable, post-treatment variables)

What Can Go Wrong and What to Do

- Risk: Inference is learning about facts we don't have with facts we have; we assume the 2 are related!
- Robustness: Imputation and analysis separated $\rightsquigarrow$ imputation affects only missing observations; High missingness reduces the property
- Vars to include: At least as much information in the imputation model as in the analysis model: all vars in analysis model; others that would help predict (e.g., All measures of a variable, post-treatment variables)
- Transform: Fit imputation model distributional assumptions by transformation to unbounded scales: $\sqrt{\text{counts}}$, $\ln(p/(1-p))$, ln(money), etc.

What Can Go Wrong and What to Do

- Risk: Inference is learning about facts we don't have with facts we have; we assume the 2 are related!
- Robustness: Imputation and analysis separated $\rightsquigarrow$ imputation affects only missing observations; High missingness reduces the property
- Vars to include: At least as much information in the imputation model as in the analysis model: all vars in analysis model; others that would help predict (e.g., All measures of a variable, post-treatment variables)
- Transform: Fit imputation model distributional assumptions by transformation to unbounded scales: $\sqrt{\text{counts}}$, $\ln(p/(1-p))$, ln(money), etc.
- Code ordinal variables: as close to interval as possible

What Can Go Wrong and What to Do (continued)

What Can Go Wrong and What to Do (continued)

- Strongly nonlinear relationships: use transformations or added quadratic terms

What Can Go Wrong and What to Do (continued)

- Strongly nonlinear relationships: use transformations or added quadratic terms
- "Uncongenial" imputations: If imputation model has as much info as analysis model, but the specification (such as the functional form) differs, CIs are conservative (e.g., $\geq 95\%$ CIs)

What Can Go Wrong and What to Do (continued)

- Strongly nonlinear relationships: use transformations or added quadratic terms
- "Uncongenial" imputations: If imputation model has as much info as analysis model, but the specification (such as the functional form) differs, CIs are conservative (e.g., $\geq 95\%$ CIs)
- Super-efficiency: When imputation model includes more information than analysis model, it can be more efficient than the "optimal" application-specific model

What Can Go Wrong and What to Do (continued)

- Strongly nonlinear relationships: use transformations or added quadratic terms
- "Uncongenial" imputations: If imputation model has as much info as analysis model, but the specification (such as the functional form) differs, CIs are conservative (e.g., ≥ 95% CIs)
- Super-efficiency: When imputation model includes more information than analysis model, it can be more efficient than the "optimal" application-specific model
- Quiz: If X is randomly imputed is there attenuation (the usual consequence of random measurement error in an explanatory variable)?

What Can Go Wrong and What to Do (continued)

- Strongly nonlinear relationships: use transformations or added quadratic terms
- "Uncongenial" imputations: If imputation model has as much info as analysis model, but the specification (such as the functional form) differs, CIs are conservative (e.g., $\geq 95\%$ CIs)
- Super-efficiency: When imputation model includes more information than analysis model, it can be more efficient than the "optimal" application-specific model
- Quiz: If X is randomly imputed is there attenuation (the usual consequence of random measurement error in an explanatory variable)?
- Quiz: If X is imputed with information from Y, is there endogeneity?

What Can Go Wrong and What to Do (continued)

- Strongly nonlinear relationships: use transformations or added quadratic terms
- "Uncongenial" imputations: If imputation model has as much info as analysis model, but the specification (such as the functional form) differs, CIs are conservative (e.g., ≥ 95% CIs)
- Super-efficiency: When imputation model includes more information than analysis model, it can be more efficient than the "optimal" application-specific model
- Quiz: If X is randomly imputed is there attenuation (the usual consequence of random measurement error in an explanatory variable)?
- Quiz: If X is imputed with information from Y, is there endogeneity?
- Answer to both: draws from the joint posterior are put back into the data; not changing the joint distribution

The Best Case for Listwise Deletion

The Best Case for Listwise Deletion

Listwise deletion is better than MI when all 4 hold:

The Best Case for Listwise Deletion

Listwise deletion is better than MI when all 4 hold:

- The analysis model is conditional on X (like regression) and functional form is correct (so listwise deletion is consistent and the characteristic robustness of regression is not lost when applied to data with slight measurement error, endogeneity, nonlinearity, etc.)

The Best Case for Listwise Deletion

Listwise deletion is better than MI when all 4 hold:

- The analysis model is conditional on X (like regression) and functional form is correct (so listwise deletion is consistent and the characteristic robustness of regression is not lost when applied to data with slight measurement error, endogeneity, nonlinearity, etc.)
- NI missingness in X and no external variables are available that could be used in an imputation stage to fix the problem

The Best Case for Listwise Deletion

Listwise deletion is better than MI when all 4 hold:

- The analysis model is conditional on X (like regression) and functional form is correct (so listwise deletion is consistent and the characteristic robustness of regression is not lost when applied to data with slight measurement error, endogeneity, nonlinearity, etc.)
- NI missingness in X and no external variables are available that could be used in an imputation stage to fix the problem
- Missingness in X is not a function of Y

The Best Case for Listwise Deletion

Listwise deletion is better than MI when all 4 hold:

- The analysis model is conditional on X (like regression) and functional form is correct (so listwise deletion is consistent and the characteristic robustness of regression is not lost when applied to data with slight measurement error, endogeneity, nonlinearity, etc.)
- NI missingness in X and no external variables are available that could be used in an imputation stage to fix the problem
- Missingness in X is not a function of Y
- n left after listwise deletion is large so that the efficiency loss does not counter balance biases induced by the other conditions

The Best Case for Listwise Deletion

Listwise deletion is better than MI when all 4 hold:

- The analysis model is conditional on X (like regression) and functional form is correct (so listwise deletion is consistent and the characteristic robustness of regression is not lost when applied to data with slight measurement error, endogeneity, nonlinearity, etc.)
- NI missingness in X and no external variables are available that could be used in an imputation stage to fix the problem
- Missingness in X is not a function of Y
- n left after listwise deletion is large so that the efficiency loss does not counter balance biases induced by the other conditions

I.e., you don't trust data to impute D_{mis} but trust it to analyze D_{obs}

Example: Support for Perot, Data Inputs

Example: Support for Perot, Data Inputs

- Research question: were voters who didn't share in economic recovery more likely to support Perot in the 1996 election?

Example: Support for Perot, Data Inputs

- Research question: were voters who didn't share in economic recovery more likely to support Perot in the 1996 election?
- Data: 1996 National Election Survey (n=1714)

Example: Support for Perot, Data Inputs

- Research question: were voters who didn't share in economic recovery more likely to support Perot in the 1996 election?
- Data: 1996 National Election Survey (n=1714)
- Dependent variable: Perot Feeling Thermometer

Example: Support for Perot, Data Inputs

- Research question: were voters who didn't share in economic recovery more likely to support Perot in the 1996 election?
- Data: 1996 National Election Survey (n=1714)
- Dependent variable: Perot Feeling Thermometer
- Key explanatory variables: retrospective and propsective evaluations of national economic performance and personal financial circumstances

Example: Support for Perot, Data Inputs

- Research question: were voters who didn't share in economic recovery more likely to support Perot in the 1996 election?
- Data: 1996 National Election Survey (n=1714)
- Dependent variable: Perot Feeling Thermometer
- Key explanatory variables: retrospective and propsective evaluations of national economic performance and personal financial circumstances
- Control variables: age, education, family income, race, gender, union membership, ideology

Example: Support for Perot, Data Inputs

- Research question: were voters who didn't share in economic recovery more likely to support Perot in the 1996 election?
- Data: 1996 National Election Survey (n=1714)
- Dependent variable: Perot Feeling Thermometer
- Key explanatory variables: retrospective and propsective evaluations of national economic performance and personal financial circumstances
- Control variables: age, education, family income, race, gender, union membership, ideology
- Extra variables for imputation model to help prediction: attention to the campaign; feeling thermometers for Clinton, Dole, Democrats, Republicans; PID; Partisan moderation; vote intention; martial status; Hispanic; party contact, number of organizations R is a paying member of, and active member of.

Example: Support for Perot, Data Inputs

- Research question: were voters who didn't share in economic recovery more likely to support Perot in the 1996 election?
- Data: 1996 National Election Survey (n=1714)
- Dependent variable: Perot Feeling Thermometer
- Key explanatory variables: retrospective and propsective evaluations of national economic performance and personal financial circumstances
- Control variables: age, education, family income, race, gender, union membership, ideology
- Extra variables for imputation model to help prediction: attention to the campaign; feeling thermometers for Clinton, Dole, Democrats, Republicans; PID; Partisan moderation; vote intention; martial status; Hispanic; party contact, number of organizations R is a paying member of, and active member of.
- Nonlinear terms: age^2

Example: Support for Perot, Data Inputs

- Research question: were voters who didn't share in economic recovery more likely to support Perot in the 1996 election?
- Data: 1996 National Election Survey (n=1714)
- Dependent variable: Perot Feeling Thermometer
- Key explanatory variables: retrospective and propsective evaluations of national economic performance and personal financial circumstances
- Control variables: age, education, family income, race, gender, union membership, ideology
- Extra variables for imputation model to help prediction: attention to the campaign; feeling thermometers for Clinton, Dole, Democrats, Republicans; PID; Partisan moderation; vote intention; martial status; Hispanic; party contact, number of organizations R is a paying member of, and active member of.
- Nonlinear terms: age^2
- Transform: to more closely approximate distributional assumptions: logged number of organizations participating in

Example: Support for Perot, Results

- Analysis model: linear regression

Example: Support for Perot, Results

- Analysis model: linear regression
- Amelia for Imputation: generate 5 imputed data sets

Example: Support for Perot, Results

- Analysis model: linear regression
- Amelia for Imputation: generate 5 imputed data sets
- Coefficient on retrospective economic evaluations (range: 1–5):

Example: Support for Perot, Results

- Analysis model: linear regression
- Amelia for Imputation: generate 5 imputed data sets
- Coefficient on retrospective economic evaluations (range: 1–5):

Listwise deletion	.43
	(.90)
Multiple imputation	1.65
	(.72)

Example: Support for Perot, Results

- Analysis model: linear regression
- Amelia for Imputation: generate 5 imputed data sets
- Coefficient on retrospective economic evaluations (range: 1–5):

Listwise deletion	.43
	(.90)
Multiple imputation	1.65
	(.72)

- Effect: $(5 - 1) \times 1.65 = 6.6$, also % of the range of Y

Example: Support for Perot, Results

- Analysis model: linear regression
- Amelia for Imputation: generate 5 imputed data sets
- Coefficient on retrospective economic evaluations (range: 1–5):

Listwise deletion	.43
	(.90)
Multiple imputation	1.65
	(.72)

- Effect: $(5 - 1) \times 1.65 = 6.6$, also % of the range of Y
- MI estimator is more efficient, with a smaller SE

Example: Support for Perot, Results

- Analysis model: linear regression
- Amelia for Imputation: generate 5 imputed data sets
- Coefficient on retrospective economic evaluations (range: 1–5):

Listwise deletion	.43
	(.90)
Multiple imputation	1.65
	(.72)

- Effect: $(5-1) \times 1.65 = 6.6$, also % of the range of Y
- MI estimator is more efficient, with a smaller SE
- MI estimator is 4 times larger

Example: Support for Perot, Results

- Analysis model: linear regression
- Amelia for Imputation: generate 5 imputed data sets
- Coefficient on retrospective economic evaluations (range: 1–5):

Listwise deletion	.43
	(.90)
Multiple imputation	1.65
	(.72)

- Effect: $(5-1) \times 1.65 = 6.6$, also % of the range of Y
- MI estimator is more efficient, with a smaller SE
- MI estimator is 4 times larger
- Based on listwise deletion: no evidence that perception of poor economic performance is related to support for Perot

Example: Support for Perot, Results

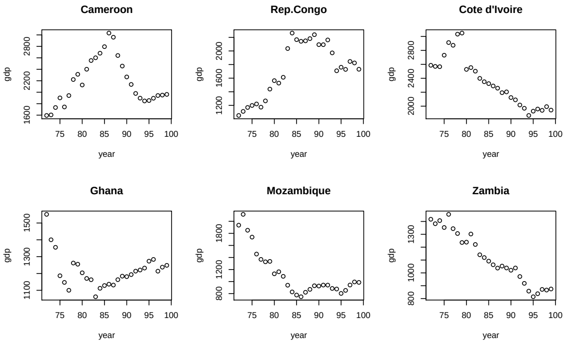
- Analysis model: linear regression
- Amelia for Imputation: generate 5 imputed data sets
- Coefficient on retrospective economic evaluations (range: 1–5):

Listwise deletion	.43
	(.90)
Multiple imputation	1.65
	(.72)

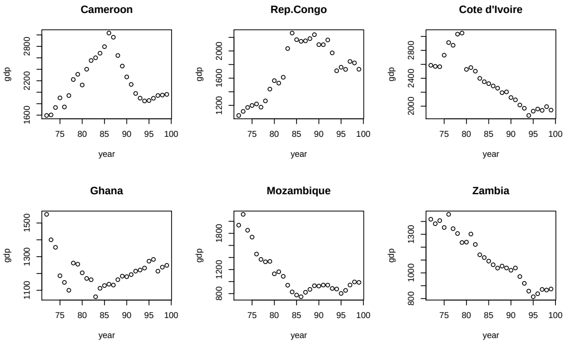
- Effect: $(5-1) \times 1.65 = 6.6$, also % of the range of Y
- MI estimator is more efficient, with a smaller SE
- MI estimator is 4 times larger
- Based on listwise deletion: no evidence that perception of poor economic performance is related to support for Perot
- Based on MI estimator: R's with negative retrospective economic evaluations more likely to have favorable views of Perot

MI in Time Series Cross-Section Data

MI in Time Series Cross-Section Data

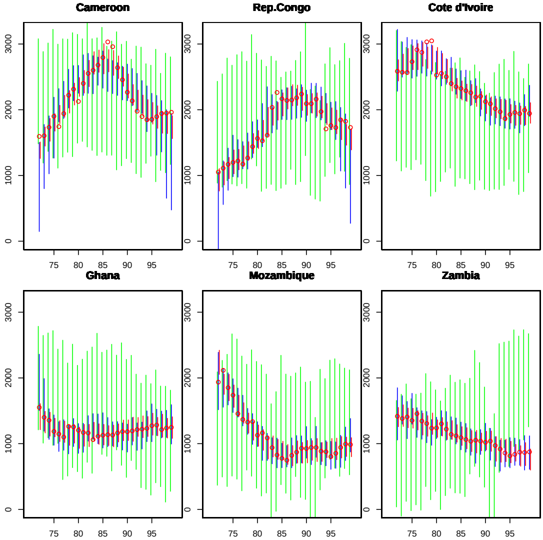


MI in Time Series Cross-Section Data



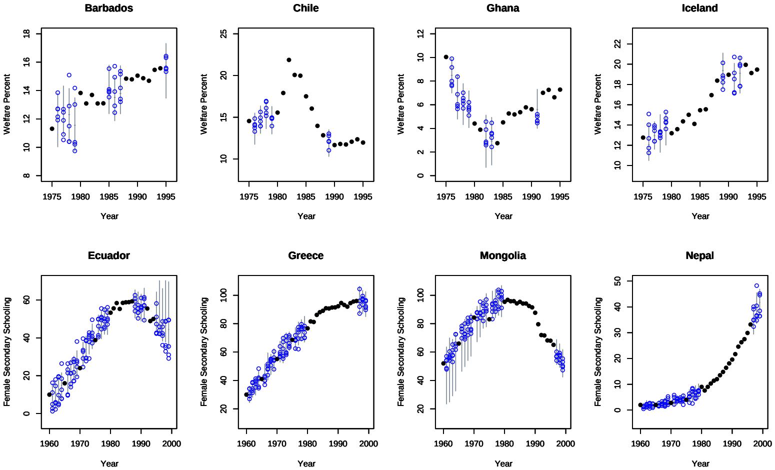
Include: (1) fixed effects, (2) time trends, and (3) priors for cells

Imputation one Observation at a time



Circles=true GDP; green=no time trends; blue=polynomials; red=LOESS

Replication of Baum and Lake: Imputation Model Fit



Black = observed. Blue circles = five imputations; Bars = 95% CIs