Advanced Quantitative Research Methodology, Lecture Notes: Missing Data[1]

Gary King

GaryKing.org

April 26, 2013

Readings

Readings

- King, Gary; James Honaker; Ann Joseph; Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *APSR* 95, 1 (March, 2001): 49-69.

Readings

- King, Gary; James Honaker; Ann Joseph; Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *APSR* 95, 1 (March, 2001): 49-69.
- James Honaker and Gary King. "What to do about Missing Values in Time Series Cross-Section Data," *AJPS* 54, 2 (April, 2010): 561-581

Readings

- King, Gary; James Honaker; Ann Joseph; Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *APSR* 95, 1 (March, 2001): 49-69.
- James Honaker and Gary King. "What to do about Missing Values in Time Series Cross-Section Data," *AJPS* 54, 2 (April, 2010): 561-581
- Blackwell, Matthew, James Honaker, and Gary King. 2013. "Multiple Overimputation: A Unified Approach to Measurement Error and Missing Data," http://j.mp/jqdj72.

Readings

- King, Gary; James Honaker; Ann Joseph; Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *APSR* 95, 1 (March, 2001): 49-69.
- James Honaker and Gary King. "What to do about Missing Values in Time Series Cross-Section Data," *AJPS* 54, 2 (April, 2010): 561-581
- Blackwell, Matthew, James Honaker, and Gary King. 2013. "Multiple Overimputation: A Unified Approach to Measurement Error and Missing Data," http://j.mp/jqdj72.
- Amelia II: A Program for Missing Data

Readings

- King, Gary; James Honaker; Ann Joseph; Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *APSR* 95, 1 (March, 2001): 49-69.
- James Honaker and Gary King. "What to do about Missing Values in Time Series Cross-Section Data," *AJPS* 54, 2 (April, 2010): 561-581
- Blackwell, Matthew, James Honaker, and Gary King. 2013. "Multiple Overimputation: A Unified Approach to Measurement Error and Missing Data," http://j.mp/jqdj72.
- Amelia II: A Program for Missing Data
- http://gking.harvard.edu/amelia

Summary

Summary

- Some common but biased or inefficient missing data practices:

Summary

- Some common but biased or inefficient missing data practices:
  - Make up numbers: e.g., changing Party ID "don't knows" to "independent"

Summary

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

Summary

- Some common but biased or inefficient missing data practices:
  - Make up numbers: e.g., changing Party ID "don't knows" to "independent"
  - Listwise deletion: used by 94% pre-2000 in AJPS/APSR/BJPS
  - Various other ad hoc approaches

Summary

- Some common but biased or inefficient missing data practices:
  - Make up numbers: e.g., changing Party ID “don't knows” to “independent”
  - Listwise deletion: used by 94% pre-2000 in AJPS/APSR/BJPS
  - Various other ad hoc approaches
- Application-specific methods: efficient, but model-dependent and hard to develop and use

Summary

- Some common but biased or inefficient missing data practices:
  - Make up numbers: e.g., changing Party ID "don't knows" to "independent"
  - Listwise deletion: used by 94% pre-2000 in AJPS/APSR/BJPS
  - Various other ad hoc approaches
- Application-specific methods: efficient, but model-dependent and hard to develop and use
- An easy-to-use and statistically appropriate alternative, Multiple imputation:

Summary

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

Summary

- Some common but biased or inefficient missing data practices:
  - Make up numbers: e.g., changing Party ID "don't knows" to "independent"
  - Listwise deletion: used by 94% pre-2000 in AJPS/APSR/BJPS
  - Various other ad hoc approaches
- Application-specific methods: efficient, but model-dependent and hard to develop and use
- An easy-to-use and statistically appropriate alternative, Multiple imputation:
  - fill in five data sets with different imputations for missing values
  - analyze each one as you would without missingness

Summary

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

Missingness Notation

Missingness Notation

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

Missingness Notation

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

D_{mis} = *missing* elements in D (in Red)

Missingness Notation

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

D_{mis} = *missing* elements in D (in Red)
D_{obs} = observed elements in D

Missingness Notation

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

D_{mis} = *missing* elements in D (in Red)
D_{obs} = observed elements in D

$\rightsquigarrow$ Missing elements must exist (what's your view on the National Helium Reserve?)

Possible Assumptions

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

Possible Assumptions

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

- Reasons for the odd terminology are historical.

Missingness Assumptions, again

Missingness Assumptions, again

1. MCAR: Coin flips determine whether to answer survey questions (naive)

Missingness Assumptions, again

1. MCAR: Coin flips determine whether to answer survey questions (naive)

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

Missingness Assumptions, again

1. MCAR: Coin flips determine whether to answer survey questions (naive)

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

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

Missingness Assumptions, again

1. MCAR: Coin flips determine whether to answer survey questions (naive)

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

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

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

Missingness Assumptions, again

1. MCAR: Coin flips determine whether to answer survey questions (naive)

$$\mathrm{P}(M|D) = \mathrm{P}(M)$$

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

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

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

Missingness Assumptions, again

1. MCAR: Coin flips determine whether to answer survey questions (naive)

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

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

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

- 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: Coin flips determine whether to answer survey questions (naive)

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

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

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

   - 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: missingness depends on unobservables (fatalistic)

Missingness Assumptions, again

1. MCAR: Coin flips determine whether to answer survey questions (naive)

$$\mathrm{P}(M|D) = \mathrm{P}(M)$$

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

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

   - 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: missingness depends on unobservables (fatalistic)
   - $P(M|D)$ doesn't simplify

Missingness Assumptions, again

1. MCAR: Coin flips determine whether to answer survey questions (naive)

$$\mathrm{P}(M|D) = \mathrm{P}(M)$$

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

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

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

Missingness Assumptions, again

1. MCAR: Coin flips determine whether to answer survey questions (naive)

$$\mathrm{P}(M|D) = \mathrm{P}(M)$$

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

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

   - 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: missingness depends on unobservables (fatalistic)
   - $P(M|D)$ 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

How Bad Is Listwise Deletion?

⏩ Skip

How Bad Is Listwise Deletion?

Skip

Goal: estimate β_1, where X_2 has λ missing values (y, X_1 are fully observed).

How Bad Is Listwise Deletion?

⏩ Skip

Goal: estimate β_1, where X_2 has λ missing values (y, X_1 are fully observed).

$$E(y) = X_1\beta_1 + X_2\beta_2$$

Goal: estimate β_1, where X_2 has λ missing values (y, X_1 are fully observed).

$$E(y) = X_1\beta_1 + X_2\beta_2$$

The choice in real research:

Goal: estimate β_1, where X_2 has λ missing values (y, X_1 are fully observed).

$$E(y) = X_1\beta_1 + X_2\beta_2$$

The choice in real research:

Infeasible Estimator Regress y on X_1 and a fully observed X_2, and use b_1^I, the coefficient on X_1.

How Bad Is Listwise Deletion?

Skip

Goal: estimate β_1, where X_2 has λ missing values (y, X_1 are fully observed).

$$E(y) = X_1\beta_1 + X_2\beta_2$$

The choice in real research:

Infeasible Estimator Regress y on X_1 and a fully observed X_2, and use b_1^I, the coefficient on X_1.

Omitted Variable Estimator Omit X_2 and estimate β_1 by b_1^O, the slope from regressing y on X_1.

How Bad Is Listwise Deletion?

Skip

Goal: estimate β_1, where X_2 has λ missing values (y, X_1 are fully observed).

$$E(y) = X_1\beta_1 + X_2\beta_2$$

The choice in real research:

Infeasible Estimator Regress y on X_1 and a fully observed X_2, and use b_1^I, the coefficient on X_1.

Omitted Variable Estimator Omit X_2 and estimate β_1 by b_1^O, the slope from regressing y on X_1.

Listwise Deletion Estimator Perform listwise deletion on $\{y, X_1, X_2\}$, and then estimate β_1 as b_1^L, the coefficient on X_1 when regressing y on X_1 and X_2.

In the *best* case scenerio for listwise deletion (MCAR), should we delete listwise or omit the variable?

Mean Square Error as a measure of the badness of an estimator $\hat{a}$ of a.

In the *best* case scenerio for listwise deletion (MCAR), should we delete listwise or omit the variable?

Mean Square Error as a measure of the badness of an estimator $\hat{a}$ of a.

$$\text{MSE}(\hat{a}) = E[(\hat{a} - a)^2]$$

In the *best* case scenerio for listwise deletion (MCAR), should we delete listwise or omit the variable?

Mean Square Error as a measure of the badness of an estimator $\hat{a}$ of a.

$$\begin{aligned}\text{MSE}(\hat{a}) &= E[(\hat{a} - a)^2] \\ &= V(\hat{a}) + [E(\hat{a} - a)]^2\end{aligned}$$

In the *best* case scenerio for listwise deletion (MCAR), should we delete listwise or omit the variable?

Mean Square Error as a measure of the badness of an estimator $\hat{a}$ of a.

$$\begin{aligned}
\text{MSE}(\hat{a}) &= E[(\hat{a} - a)^2] \\
&= V(\hat{a}) + [E(\hat{a} - a)]^2 \\
&= \text{Variance}(\hat{a}) + \text{bias}(\hat{a})^2
\end{aligned}$$

In the *best* case scenerio for listwise deletion (MCAR), should we delete listwise or omit the variable?

Mean Square Error as a measure of the badness of an estimator $\hat{a}$ of a.

$$
\begin{aligned}
\text{MSE}(\hat{a}) &= E[(\hat{a} - a)^2] \\
&= V(\hat{a}) + [E(\hat{a} - a)]^2 \\
&= \text{Variance}(\hat{a}) + \text{bias}(\hat{a})^2
\end{aligned}
$$

To compare, compute

In the *best* case scenerio for listwise deletion (MCAR), should we delete listwise or omit the variable?

Mean Square Error as a measure of the badness of an estimator $\hat{a}$ of a.

$$\begin{aligned}
\text{MSE}(\hat{a}) &= E[(\hat{a} - a)^2] \\
&= V(\hat{a}) + [E(\hat{a} - a)]^2 \\
&= \text{Variance}(\hat{a}) + \text{bias}(\hat{a})^2
\end{aligned}$$

To compare, compute

$$\text{M}SE(b_1^L) - \text{MSE}(b_1^O) =$$

In the *best* case scenerio for listwise deletion (MCAR), should we delete listwise or omit the variable?

Mean Square Error as a measure of the badness of an estimator $\hat{a}$ of a.

$$\begin{aligned}\text{MSE}(\hat{a}) &= E[(\hat{a}-a)^2]\\ &= V(\hat{a}) + [E(\hat{a}-a)]^2\\ &= \text{Variance}(\hat{a}) + \text{bias}(\hat{a})^2\end{aligned}$$

To compare, compute

$$\text{M}SE(b_1^L) - \text{MSE}(b_1^O) = \begin{cases} > 0 & \text{when omitting the variable is better}\\ < 0 & \text{when listwise deletion is better}\end{cases}$$

Derivation and Implications

Derivation and Implications

$$\text{MSE}(b_1^L) - \text{MSE}(b_1^O)$$

Derivation and Implications

$$\text{MSE}(b_1^L) - \text{MSE}(b_1^O) = \left(\frac{\lambda}{1-\lambda}\text{V}(b_1^I)\right) + F[\text{V}(b_2^I) - \beta_2\beta_2']F'$$

Derivation and Implications

$$
\begin{aligned}
\text{MSE}(b_1^L) - \text{MSE}(b_1^O) &= \left(\frac{\lambda}{1-\lambda}\text{V}(b_1^I)\right) + F[\text{V}(b_2^I) - \beta_2\beta_2']F' \\
&= (\text{Missingness part}) + (\text{Observed part})
\end{aligned}
$$

Derivation and Implications

$$\text{MSE}(b_1^L) - \text{MSE}(b_1^O) = \left(\frac{\lambda}{1-\lambda}\text{V}(b_1^I)\right) + F[\text{V}(b_2^I) - \beta_2\beta_2']F'$$
$$= (\text{Missingness part}) + (\text{Observed part})$$

1. Missingness part (> 0) is an extra tilt away from listwise deletion

Derivation and Implications

$$\mathrm{MSE}(b_1^L) - \mathrm{MSE}(b_1^O) = \left(\frac{\lambda}{1-\lambda}\mathrm{V}(b_1^I)\right) + F[\mathrm{V}(b_2^I) - \beta_2\beta_2']F'$$
$$= (\text{Missingness part}) + (\text{Observed part})$$

1. Missingness part (> 0) is an extra tilt away from listwise deletion
2. Observed part is the standard bias-efficiency tradeoff of omitting variables, even without missingness

Derivation and Implications

$$\text{MSE}(b_1^L) - \text{MSE}(b_1^O) = \left(\frac{\lambda}{1-\lambda}\text{V}(b_1^I)\right) + F[\text{V}(b_2^I) - \beta_2\beta_2']F'$$
$$= (\text{Missingness part}) + (\text{Observed part})$$

1. Missingness part (> 0) is an extra tilt away from listwise deletion
2. Observed part is the standard bias-efficiency tradeoff of omitting variables, even without missingness
3. How big is λ usually?

Derivation and Implications

$$\text{MSE}(b_1^L) - \text{MSE}(b_1^O) = \left(\frac{\lambda}{1-\lambda}\text{V}(b_1^I)\right) + F[\text{V}(b_2^I) - \beta_2\beta_2']F'$$
$$= (\text{Missingness part}) + (\text{Observed part})$$

1. Missingness part (> 0) is an extra tilt away from listwise deletion
2. Observed part is the standard bias-efficiency tradeoff of omitting variables, even without missingness
3. How big is λ usually?
   - $\lambda \approx 1/3$ on average in real political science articles

Derivation and Implications

$$\text{MSE}(b_1^L) - \text{MSE}(b_1^O) = \left(\frac{\lambda}{1-\lambda}\text{V}(b_1^I)\right) + F[\text{V}(b_2^I) - \beta_2\beta_2']F'$$
$$= (\text{Missingness part}) + (\text{Observed part})$$

1. Missingness part (> 0) is an extra tilt away from listwise deletion
2. Observed part is the standard bias-efficiency tradeoff of omitting variables, even without missingness
3. How big is λ usually?
    - $\lambda \approx 1/3$ on average in real political science articles
    - $> 1/2$ at a recent SPM Conference

Derivation and Implications

$$
\begin{aligned}
\text{MSE}(b_1^L) - \text{MSE}(b_1^O) &= \left(\frac{\lambda}{1-\lambda}\text{V}(b_1^I)\right) + F[\text{V}(b_2^I) - \beta_2\beta_2']F' \\
&= (\text{Missingness part}) + (\text{Observed part})
\end{aligned}
$$

1. Missingness part (> 0) is an extra tilt away from listwise deletion
2. Observed part is the standard bias-efficiency tradeoff of omitting variables, even without missingness
3. How big is λ usually?
   - $\lambda \approx 1/3$ on average in real political science articles
   - $> 1/2$ at a recent SPM Conference
   - Larger for authors who work harder to avoid omitted variable bias

Derivation and Implications

4. If $\lambda \approx 0.5$, the contribution of the missingness (tilting away from choosing listwise deletion over omitting variables) is

Derivation and Implications

4. If $\lambda \approx 0.5$, the contribution of the missingness (tilting away from choosing listwise deletion over omitting variables) is

$$\text{RMSE difference} = \sqrt{\frac{\lambda}{1-\lambda} V(b_1^I)} = \sqrt{\frac{0.5}{1-0.5}} \text{SE}(b_1^I) = \text{SE}(b_1^I)$$

Derivation and Implications

4. If $\lambda \approx 0.5$, the contribution of the missingness (tilting away from choosing listwise deletion over omitting variables) is

$$\text{RMSE difference} = \sqrt{\frac{\lambda}{1-\lambda} V(b_1^I)} = \sqrt{\frac{0.5}{1-0.5}} \text{SE}(b_1^I) = \text{SE}(b_1^I)$$

(The sqrt of only one piece, for simplicity, not the difference.)

Derivation and Implications

4. If $\lambda \approx 0.5$, the contribution of the missingness (tilting away from choosing listwise deletion over omitting variables) is

$$\text{RMSE difference} = \sqrt{\frac{\lambda}{1-\lambda} V(b_1^I)} = \sqrt{\frac{0.5}{1-0.5}} \text{SE}(b_1^I) = \text{SE}(b_1^I)$$

   (The sqrt of only one piece, for simplicity, not the difference.)

5. **Result:** The point estimate in the average political science article is about an additional standard error farther away from the truth because of listwise deletion (as compared to omitting X_2 entirely).

Derivation and Implications

4. If $\lambda \approx 0.5$, the contribution of the missingness (tilting away from choosing listwise deletion over omitting variables) is

$$\text{RMSE difference} = \sqrt{\frac{\lambda}{1-\lambda} V(b_1^I)} = \sqrt{\frac{0.5}{1-0.5}} \text{SE}(b_1^I) = \text{SE}(b_1^I)$$

(The sqrt of only one piece, for simplicity, not the difference.)

5. **Result:** The point estimate in the average political science article is about an additional standard error farther away from the truth because of listwise deletion (as compared to omitting X_2 entirely).
6. **Conclusion**: Listwise deletion is often as bad a problem as the much better known omitted variable bias — in the best case scenerio (MCAR)

Existing General Purpose Missing Data Methods

◂ Return

Existing General Purpose Missing Data Methods

Return

Fill in or delete the missing data, and then act as if there were no missing data. None work in general under MAR.

Existing General Purpose Missing Data Methods

◂ Return

Fill in or delete the missing data, and then act as if there were no missing data. None work in general under MAR.

1. Listwise deletion (RMSE is 1 SE off if MCAR holds; biased under MAR)

Existing General Purpose Missing Data Methods

◂ Return

Fill in or delete the missing data, and then act as if there were no missing data. None work in general under MAR.

1. Listwise deletion (RMSE is 1 SE off if MCAR holds; biased under MAR)
2. Best guess imputation (depends on guesser!)

Existing General Purpose Missing Data Methods

Return

Fill in or delete the missing data, and then act as if there were no missing data. None work in general under MAR.

1. Listwise deletion (RMSE is 1 SE off if MCAR holds; biased under MAR)
2. Best guess imputation (depends on guesser!)
3. Imputing a zero and then adding an additional dummy variable to control for the imputed value (biased)

Existing General Purpose Missing Data Methods

◂ Return

Fill in or delete the missing data, and then act as if there were no missing data. None work in general under MAR.

1. Listwise deletion (RMSE is 1 SE off if MCAR holds; biased under MAR)
2. Best guess imputation (depends on guesser!)
3. Imputing a zero and then adding an additional dummy variable to control for the imputed value (biased)
4. Pairwise deletion (assumes MCAR)

Existing General Purpose Missing Data Methods

Return

Fill in or delete the missing data, and then act as if there were no missing data. None work in general under MAR.

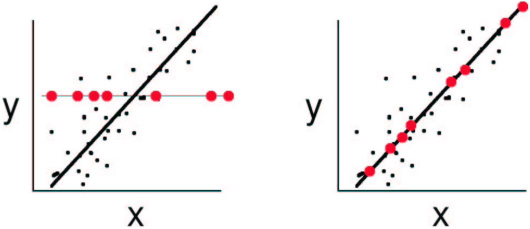
1. Listwise deletion (RMSE is 1 SE off if MCAR holds; biased under MAR)
2. Best guess imputation (depends on guesser!)
3. Imputing a zero and then adding an additional dummy variable to control for the imputed value (biased)
4. Pairwise deletion (assumes MCAR)
5. Hot deck imputation, (inefficient, standard errors wrong)

Existing General Purpose Missing Data Methods

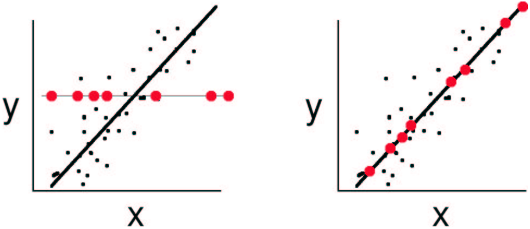
Return

Fill in or delete the missing data, and then act as if there were no missing data. None work in general under MAR.

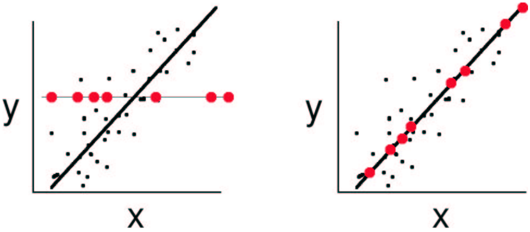
1. Listwise deletion (RMSE is 1 SE off if MCAR holds; biased under MAR)
2. Best guess imputation (depends on guesser!)
3. Imputing a zero and then adding an additional dummy variable to control for the imputed value (biased)
4. Pairwise deletion (assumes MCAR)
5. Hot deck imputation, (inefficient, standard errors wrong)
6. Mean substitution (attenuates estimated relationships)



y
x
y
x



7. y-hat regression imputation, (optimistic: scatter when observed, perfectly linear when unobserved; SEs too small)



7. y-hat regression imputation, (optimistic: scatter when observed, perfectly linear when unobserved; SEs too small)
8. y-hat + ϵ regression imputation (assumes no estimation uncertainty, does not help for scattered missingness)

Application-specific Methods for Missing Data

Application-specific Methods for Missing Data

1. Base inferences on the likelihood function or posterior distribution, by conditioning on observed data only, $P(\theta|Y_{obs})$.

Application-specific Methods for Missing Data

1. Base inferences on the likelihood function or posterior distribution, by conditioning on observed data only, $P(\theta | Y_{obs})$.
2. E.g., models of censoring, truncation, etc.

Application-specific Methods for Missing Data

1. Base inferences on the likelihood function or posterior distribution, by conditioning on observed data only, $P(\theta|Y_{obs})$.
2. E.g., models of censoring, truncation, etc.
3. Optimal theoretically, if specification is correct

Application-specific Methods for Missing Data

1. Base inferences on the likelihood function or posterior distribution, by conditioning on observed data only, $P(\theta|Y_{obs})$.
2. E.g., models of censoring, truncation, etc.
3. Optimal theoretically, if specification is correct
4. Not robust (i.e., sensitive to distributional assumptions) if model is incorrect

Application-specific Methods for Missing Data

1. Base inferences on the likelihood function or posterior distribution, by conditioning on observed data only, $P(\theta|Y_{obs})$.
2. E.g., models of censoring, truncation, etc.
3. Optimal theoretically, if specification is correct
4. Not robust (i.e., sensitive to distributional assumptions) if model is incorrect
5. Often difficult practically

Application-specific Methods for Missing Data

1. Base inferences on the likelihood function or posterior distribution, by conditioning on observed data only, $P(\theta|Y_{obs})$.
2. E.g., models of censoring, truncation, etc.
3. Optimal theoretically, if specification is correct
4. Not robust (i.e., sensitive to distributional assumptions) if model is incorrect
5. Often difficult practically
6. Very difficult with missingness scattered through X and Y

How to create application-specific methods?

How to create application-specific methods?

Skip

1. We observe M always. Suppose we also see all the contents of D.

How to create application-specific methods?

Skip

1. We observe M always. Suppose we also see all the contents of D.
2. Then the likelihood is

1. We observe M always. Suppose we also see all the contents of D.
2. Then the likelihood is

$$P(D, M|\theta, \gamma) = P(D|\theta)P(M|D, \gamma),$$

1. We observe M always. Suppose we also see all the contents of D.
2. Then the likelihood is

$$P(D, M|\theta, \gamma) = P(D|\theta)P(M|D, \gamma),$$

the likelihood if D were observed, and the model for missingness.

How to create application-specific methods?

▸▸ Skip

1. We observe M always. Suppose we also see all the contents of D.
2. Then the likelihood is

$$P(D, M|\theta, \gamma) = P(D|\theta)P(M|D, \gamma),$$

   the likelihood if D were observed, and the model for missingness.

   - If D and M are observed, when can we drop $P(M|D, \gamma)$?

How to create application-specific methods?

▸▸ Skip

1. We observe M always. Suppose we also see all the contents of D.
2. Then the likelihood is

$$P(D, M|\theta, \gamma) = P(D|\theta)P(M|D, \gamma),$$

the likelihood if D were observed, and the model for missingness.

- If D and M are observed, when can we drop $P(M|D, \gamma)$?
- Stochastic and parametric independence

How to create application-specific methods?

▸▸ Skip

1. We observe M always. Suppose we also see all the contents of D.
2. Then the likelihood is

$$P(D, M|\theta, \gamma) = P(D|\theta)P(M|D, \gamma),$$

   the likelihood if D were observed, and the model for missingness.

   - If D and M are observed, when can we drop $P(M|D, \gamma)$?
   - Stochastic and parametric independence

3. Suppose now D is observed (as usual) only when M is 1.

4. Then the likelihood integrates out the missing observations

4. Then the likelihood integrates out the missing observations

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

4. Then the likelihood integrates out the missing observations

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

4. Then the likelihood integrates out the missing observations

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

and if assume MAR (D and M are stochastically and parametrically independent), then

4. Then the likelihood integrates out the missing observations

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

and if assume MAR (D and M are stochastically and parametrically independent), then

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

4. Then the likelihood integrates out the missing observations

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

and if assume MAR (D and M are stochastically and parametrically independent), then

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

4. Then the likelihood integrates out the missing observations

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

and if assume MAR (D and M are stochastically and parametrically independent), then

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

because $P(M|D_{obs}, \gamma)$ is constant w.r.t. θ

4. Then the likelihood integrates out the missing observations

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

and if assume MAR (D and M are stochastically and parametrically independent), then

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

because $P(M|D_{obs}, \gamma)$ is constant w.r.t. θ

5. Without the MAR assumption, the missingness model can't be dropped; its an NI model.

4. Then the likelihood integrates out the missing observations

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

and if assume MAR (D and M are stochastically and parametrically independent), then

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

because $P(M|D_{obs}, \gamma)$ is constant w.r.t. θ

5. Without the MAR assumption, the missingness model can't be dropped; its an NI model.
6. Specifying the missingness mechanism is hard. Little theory is available

4. Then the likelihood integrates out the missing observations

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

and if assume MAR (D and M are stochastically and parametrically independent), then

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

because $P(M|D_{obs}, \gamma)$ is constant w.r.t. θ

5. Without the MAR assumption, the missingness model can't be dropped; its an NI model.
6. Specifying the missingness mechanism is hard. Little theory is available
7. NI models (Heckman, many others) haven't always done well when truth is known

Multiple Imputation

Return

Point estimates are consistent, efficient, and the standard errors are right. To compute:

Multiple Imputation

Return

Point estimates are consistent, efficient, and the standard errors are right. To compute:

1. Impute m values for each missing element

Multiple Imputation

Return

Point estimates are consistent, efficient, and the standard errors are right. To compute:

1. Impute m values for each missing element
   (a) Imputation method assumes MAR

Multiple Imputation

◂ Return

Point estimates are consistent, efficient, and the standard errors are right. To compute:

1. Impute m values for each missing element
   (a) Imputation method assumes MAR
   (b) Uses a model with stochastic and systematic components

Point estimates are consistent, efficient, and the standard errors are right. To compute:

1. Impute m values for each missing element
   - (a) Imputation method assumes MAR
   - (b) Uses a model with stochastic and systematic components
   - (c) Produces independent imputations

Multiple Imputation

Return

Point estimates are consistent, efficient, and the standard errors are right. To compute:

1. Impute m values for each missing element
   (a) Imputation method assumes MAR
   (b) Uses a model with stochastic and systematic components
   (c) Produces independent imputations
   (d) (We'll give you a model to impute later)

Multiple Imputation

◂ Return

Point estimates are consistent, efficient, and the standard errors are right.
To compute:

1. Impute *m* values for each missing element
    (a) Imputation method assumes MAR
    (b) Uses a model with stochastic and systematic components
    (c) Produces independent imputations
    (d) (We'll give you a model to impute later)
2. Create *m* completed data sets

Multiple Imputation

Return

Point estimates are consistent, efficient, and the standard errors are right. To compute:

1. Impute m values for each missing element
   (a) Imputation method assumes MAR
   (b) Uses a model with stochastic and systematic components
   (c) Produces independent imputations
   (d) (We'll give you a model to impute later)
2. Create m completed data sets
   (a) Observed data are the same across the data sets

Multiple Imputation

◂ Return

Point estimates are consistent, efficient, and the standard errors are right.

To compute:

1. Impute m values for each missing element
   (a) Imputation method assumes MAR
   (b) Uses a model with stochastic and systematic components
   (c) Produces independent imputations
   (d) (We'll give you a model to impute later)
2. Create m completed data sets
   (a) Observed data are the same across the data sets
   (b) Imputations of missing data differ

Multiple Imputation

Return

Point estimates are consistent, efficient, and the standard errors are right.
To compute:

1. Impute *m* values for each missing element
   (a) Imputation method assumes MAR
   (b) Uses a model with stochastic and systematic components
   (c) Produces independent imputations
   (d) (We'll give you a model to impute later)
2. Create *m* completed data sets
   (a) Observed data are the same across the data sets
   (b) Imputations of missing data differ
      i. Cells we can predict well don't differ much

Multiple Imputation

Return

Point estimates are consistent, efficient, and the standard errors are right.
To compute:

1. Impute m values for each missing element
   (a) Imputation method assumes MAR
   (b) Uses a model with stochastic and systematic components
   (c) Produces independent imputations
   (d) (We'll give you a model to impute later)
2. Create m completed data sets
   (a) Observed data are the same across the data sets
   (b) Imputations of missing data differ
      i. Cells we can predict well don't differ much
      ii. Cells we can't predict well differ a lot

Point estimates are consistent, efficient, and the standard errors are right. To compute:

1. Impute m values for each missing element
   (a) Imputation method assumes MAR
   (b) Uses a model with stochastic and systematic components
   (c) Produces independent imputations
   (d) (We'll give you a model to impute later)
2. Create m completed data sets
   (a) Observed data are the same across the data sets
   (b) Imputations of missing data differ
      i. Cells we can predict well don't differ much
      ii. Cells we can't predict well differ a lot
3. Run whatever statistical method you would have with no missing data for each completed data set

4. Overall Point estimate: average individual point estimates, q_j $(j = 1, \ldots, m)$:

4. Overall Point estimate: average individual point estimates, q_j $(j = 1, \ldots, m)$:

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

4. Overall 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: use this equation:

4. Overall 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: use this equation:

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

4. Overall 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: use this equation:

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

4. Overall 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: use this equation:

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

Last piece vanishes as m increases

4. Overall 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: use this equation:

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

Last piece vanishes as m increases

5. Easier by simulation: draw $1/m$ sims from each data set of the QOI, combine (i.e., concatenate into a larger set of simulations), and make inferences as usual.

A General Model for Imputations

A General Model for Imputations

1. If data were complete, we could use:

A General Model for Imputations

1. If data were complete, we could use:

$$L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma)$$

A General Model for Imputations

1. If data were complete, we could use:

$$L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma) \qquad \text{(a SURM model without } X\text{)}$$

A General Model for Imputations

1. If data were complete, we could use:

$$L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma) \qquad \text{(a SURM model without } X\text{)}$$

2. With missing data, this becomes:

A General Model for Imputations

1. If data were complete, we could use:

$$L(\mu, \Sigma|D) \propto \prod_{i=1}^{n} N(D_i|\mu, \Sigma) \qquad \text{(a SURM model without } X\text{)}$$

2. With missing data, this becomes:

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

A General Model for Imputations

1. If data were complete, we could use:

$$L(\mu, \Sigma|D) \propto \prod_{i=1}^{n} N(D_i|\mu, \Sigma) \qquad \text{(a SURM model without } X\text{)}$$

2. With missing data, this becomes:

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

A General Model for Imputations

1. If data were complete, we could use:

$$L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma) \qquad \text{(a SURM model without } X\text{)}$$

2. With missing data, this becomes:

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

   since marginals of MVN's are normal.

A General Model for Imputations

1. If data were complete, we could use:

$$L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma) \qquad \text{(a SURM model without } X\text{)}$$

2. With missing data, this becomes:

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

   since marginals of MVN's are normal.

3. Simple theoretically: merely a likelihood model for data (D_{obs}, M) and same parameters as when fully observed (μ, Σ).

A General Model for Imputations

1. If data were complete, we could use:

$$L(\mu, \Sigma|D) \propto \prod_{i=1}^{n} N(D_i|\mu, \Sigma) \qquad \text{(a SURM model without } X\text{)}$$

2. With missing data, this becomes:

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

   since marginals of MVN's are normal.

3. Simple theoretically: merely a likelihood model for data (D_{obs}, M) and same parameters as when fully observed (μ, Σ).
4. Difficult computationally: $D_{i,obs}$ has different elements observed for each i and so each observation is informative about different pieces of (μ, Σ).

5. Difficult Statistically: number of parameters increases quickly in the number of variables (p, columns of D):

5. Difficult Statistically: number of parameters increases quickly in the number of variables (p, columns of D):

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

5. Difficult Statistically: number of parameters increases quickly in the number of variables (p, columns of D):

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

5. Difficult Statistically: number of parameters increases quickly in the number of variables (p, columns of D):

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

5. Difficult Statistically: number of parameters increases quickly in the number of variables (p, columns of D):

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

E.g., for $p = 5$, parameters$= 20$; for $p = 40$ parameters$= 860$ (Compare to n.)

5. Difficult Statistically: number of parameters increases quickly in the number of variables (p, columns of D):

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

   E.g., for $p = 5$, parameters$= 20$; for $p = 40$ parameters$= 860$ (Compare to n.)

6. More appropriate models, such as for categorical or mixed variables, are harder to apply and do not usually perform better than this model (If you're going to use a difficult imputation method, you might as well use an application-specific method. Our goal is an easy-to-apply, generally applicable, method even if 2nd best.)

5. Difficult Statistically: number of parameters increases quickly in the number of variables (p, columns of D):

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

   E.g., for $p = 5$, parameters$= 20$; for $p = 40$ parameters$= 860$ (Compare to n.)

6. More appropriate models, such as for categorical or mixed variables, are harder to apply and do not usually perform better than this model (If you're going to use a difficult imputation method, you might as well use an application-specific method. Our goal is an easy-to-apply, generally applicable, method even if 2nd best.)

7. For social science survey data, which mostly contain ordinal scales, this is a reasonable choice for imputation, even though it may not be a good choice for analysis.

How to create imputations from this model

How to create imputations from this model

1. E.g., suppose D has only 2 variables, $D = \{X, Y\}$

How to create imputations from this model

1. E.g., suppose D has only 2 variables, $D = \{X, Y\}$
2. X is fully observed, Y has some missingness.

How to create imputations from this model

1. E.g., suppose D has only 2 variables, $D = \{X, Y\}$
2. X is fully observed, Y has some missingness.
3. Then $D = \{Y, X\}$ is bivariate normal:

How to create imputations from this model

1. E.g., suppose D has only 2 variables, $D = \{X, Y\}$
2. X is fully observed, Y has some missingness.
3. Then $D = \{Y, X\}$ is bivariate normal:

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

How to create imputations from this model

1. E.g., suppose D has only 2 variables, $D = \{X, Y\}$
2. X is fully observed, Y has some missingness.
3. Then $D = \{Y, X\}$ is bivariate normal:

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

How to create imputations from this model

1. E.g., suppose D has only 2 variables, $D = \{X, Y\}$
2. X is fully observed, Y has some missingness.
3. Then $D = \{Y, X\}$ is bivariate normal:

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

4. Conditionals of bivariate normals are normal:

How to create imputations from this model

1. E.g., suppose D has only 2 variables, $D = \{X, Y\}$
2. X is fully observed, Y has some missingness.
3. Then $D = \{Y, X\}$ is bivariate normal:

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

4. Conditionals of bivariate normals are normal:

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

How to create imputations from this model

1. E.g., suppose D has only 2 variables, $D = \{X, Y\}$
2. X is fully observed, Y has some missingness.
3. Then $D = \{Y, X\}$ is bivariate normal:

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

4. Conditionals of bivariate normals are normal:

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

$$E(Y|X) = \mu_y + \beta(X - \mu_x) \qquad \text{(a linear regression!)}$$

How to create imputations from this model

1. E.g., suppose D has only 2 variables, $D = \{X, Y\}$
2. X is fully observed, Y has some missingness.
3. Then $D = \{Y, X\}$ is bivariate normal:

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

4. Conditionals of bivariate normals are normal:

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

$$\begin{aligned} E(Y|X) &= \mu_y + \beta(X - \mu_x) \qquad \text{(a linear regression!)} \\ \beta &= \sigma_{xy}/\sigma_x \end{aligned}$$

How to create imputations from this model

1. E.g., suppose D has only 2 variables, $D = \{X, Y\}$
2. X is fully observed, Y has some missingness.
3. Then $D = \{Y, X\}$ is bivariate normal:

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

4. Conditionals of bivariate normals are normal:

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

$$\begin{aligned} E(Y|X) &= \mu_y + \beta(X - \mu_x) \qquad \text{(a linear regression!)} \\ \beta &= \sigma_{xy}/\sigma_x \\ V(Y|X) &= \sigma_y - \sigma_{xy}^2/\sigma_x \end{aligned}$$

5. To create imputations:

5. To create imputations:
   (a) Estimate the posterior density of μ and Σ

5. To create imputations:
    (a) Estimate the posterior density of μ and Σ
        i. Could do the usual: maximize likelihood, assume CLT applies, and draw from the normal approximation. (Hard to do, and CLT isn't a good asymptotic approximation due to large number of parameters.)

5. To create imputations:
   (a) Estimate the posterior density of μ and Σ
      i. Could do the usual: maximize likelihood, assume CLT applies, and draw from the normal approximation. (Hard to do, and CLT isn't a good asymptotic approximation due to large number of parameters.)
      ii. We will improve on this shortly

5. To create imputations:
   (a) Estimate the posterior density of μ and Σ
      i. Could do the usual: maximize likelihood, assume CLT applies, and draw from the normal approximation. (Hard to do, and CLT isn't a good asymptotic approximation due to large number of parameters.)
      ii. We will improve on this shortly
   (b) Draw μ and Σ from their posterior density

5. To create imputations:
   (a) Estimate the posterior density of μ and Σ
      i. Could do the usual: maximize likelihood, assume CLT applies, and draw from the normal approximation. (Hard to do, and CLT isn't a good asymptotic approximation due to large number of parameters.)
      ii. We will improve on this shortly
   (b) Draw μ and Σ from their posterior density
   (c) Compute simulations of $E(Y|X)$ and $V(Y|X)$ deterministically

5. To create imputations:
    (a) Estimate the posterior density of μ and Σ
        i. Could do the usual: maximize likelihood, assume CLT applies, and draw from the normal approximation. (Hard to do, and CLT isn't a good asymptotic approximation due to large number of parameters.)
        ii. We will improve on this shortly
    (b) Draw μ and Σ from their posterior density
    (c) Compute simulations of $E(Y|X)$ and $V(Y|X)$ deterministically
    (d) Draw a simulation of the missing Y from the conditional normal

5. To create imputations:
    (a) Estimate the posterior density of μ and Σ
        i. Could do the usual: maximize likelihood, assume CLT applies, and draw from the normal approximation. (Hard to do, and CLT isn't a good asymptotic approximation due to large number of parameters.)
        ii. We will improve on this shortly
    (b) Draw μ and Σ from their posterior density
    (c) Compute simulations of $E(Y|X)$ and $V(Y|X)$ deterministically
    (d) Draw a simulation of the missing Y from the conditional normal
6. In this simple example (X fully observed), this is equivalent to simulating from a linear regression of Y on X,

5. To create imputations:
   (a) Estimate the posterior density of μ and Σ
      i. Could do the usual: maximize likelihood, assume CLT applies, and draw from the normal approximation. (Hard to do, and CLT isn't a good asymptotic approximation due to large number of parameters.)
      ii. We will improve on this shortly
   (b) Draw μ and Σ from their posterior density
   (c) Compute simulations of $E(Y|X)$ and $V(Y|X)$ deterministically
   (d) Draw a simulation of the missing Y from the conditional normal
6. In this simple example (X fully observed), this is equivalent to simulating from a linear regression of Y on X,

$$\tilde{y}_i = x_i\tilde{\beta} + \tilde{\epsilon}_i,$$

5. To create imputations:
    (a) Estimate the posterior density of μ and Σ
        i. Could do the usual: maximize likelihood, assume CLT applies, and draw from the normal approximation. (Hard to do, and CLT isn't a good asymptotic approximation due to large number of parameters.)
        ii. We will improve on this shortly
    (b) Draw μ and Σ from their posterior density
    (c) Compute simulations of $E(Y|X)$ and $V(Y|X)$ deterministically
    (d) Draw a simulation of the missing Y from the conditional normal
6. In this simple example (X fully observed), this is equivalent to simulating from a linear regression of Y on X,

$$\tilde{y}_i = x_i\tilde{\beta} + \tilde{\epsilon}_i,$$

with estimation and fundamental uncertainty

Computational Algorithms

Computational Algorithms

1. Optim with hundreds of parameters would work but slowly.

Computational Algorithms

1. Optim with hundreds of parameters would work but slowly.
2. EM (expectation maximization): another algorithm for finding the maximum.

Computational Algorithms

1. Optim with hundreds of parameters would work but slowly.
2. EM (expectation maximization): another algorithm for finding the maximum.
   (a) Much faster than optim

Computational Algorithms

1. Optim with hundreds of parameters would work but slowly.
2. EM (expectation maximization): another algorithm for finding the maximum.
   (a) Much faster than optim
   (b) Intuition:

Computational Algorithms

1. Optim with hundreds of parameters would work but slowly.
2. EM (expectation maximization): another algorithm for finding the maximum.
   (a) Much faster than optim
   (b) Intuition:
      i. Without missingness, estimating β would be easy: run LS.

Computational Algorithms

1. Optim with hundreds of parameters would work but slowly.
2. EM (expectation maximization): another algorithm for finding the maximum.
   (a) Much faster than optim
   (b) Intuition:
      i. Without missingness, estimating β would be easy: run LS.
      ii. If β were known, imputation would be easy: draw $\tilde{\epsilon}$ from normal, and use $\tilde{y} = x\beta + \tilde{\epsilon}$.

Computational Algorithms

1. Optim with hundreds of parameters would work but slowly.
2. EM (expectation maximization): another algorithm for finding the maximum.
   (a) Much faster than optim
   (b) Intuition:
      i. Without missingness, estimating β would be easy: run LS.
      ii. If β were known, imputation would be easy: draw $\tilde{\epsilon}$ from normal, and use $\tilde{y} = x\beta + \tilde{\epsilon}$.
   (c) EM works by iterating between

Computational Algorithms

1. Optim with hundreds of parameters would work but slowly.
2. EM (expectation maximization): another algorithm for finding the maximum.
   (a) Much faster than optim
   (b) Intuition:
      i. Without missingness, estimating β would be easy: run LS.
      ii. If β were known, imputation would be easy: draw $\tilde{\epsilon}$ from normal, and use $\tilde{y} = x\beta + \tilde{\epsilon}$.
   (c) EM works by iterating between
      i. Impute $\hat{Y}$ with $x\hat{\beta}$, given current estimates, $\hat{\beta}$

Computational Algorithms

1. Optim with hundreds of parameters would work but slowly.
2. EM (expectation maximization): another algorithm for finding the maximum.
   (a) Much faster than optim
   (b) Intuition:
      i. Without missingness, estimating β would be easy: run LS.
      ii. If β were known, imputation would be easy: draw $\tilde{\epsilon}$ from normal, and use $\tilde{y} = x\beta + \tilde{\epsilon}$.
   (c) EM works by iterating between
      i. Impute $\hat{Y}$ with $x\hat{\beta}$, given current estimates, $\hat{\beta}$
      ii. Estimate parameters $\hat{\beta}$ (by LS) on data filled in with current imputations for Y

Computational Algorithms

1. Optim with hundreds of parameters would work but slowly.
2. EM (expectation maximization): another algorithm for finding the maximum.
   (a) Much faster than optim
   (b) Intuition:
      i. Without missingness, estimating β would be easy: run LS.
      ii. If β were known, imputation would be easy: draw $\tilde{\epsilon}$ from normal, and use $\tilde{y} = x\beta + \tilde{\epsilon}$.
   (c) EM works by iterating between
      i. Impute $\hat{Y}$ with $x\hat{\beta}$, given current estimates, $\hat{\beta}$
      ii. Estimate parameters $\hat{\beta}$ (by LS) on data filled in with current imputations for Y
   (d) Can easily do imputation via $x\hat{\beta} + \tilde{\epsilon}$, but SEs too small due to no estimation uncertainty ($\hat{\beta} \neq \beta$); i.e., we need to draw β from its posterior first

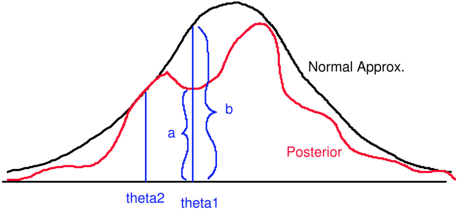
3. EMs: EM for maximization and then simulation (as usual) from asymptotic normal posterior

3. EMs: EM for maximization and then simulation (as usual) from asymptotic normal posterior
   (a) EMs adds estimation uncertainty to EM imputations by drawing $\tilde{\beta}$ from its asymptotic normal distribution, $N(\hat{\beta}, \hat{V}(\hat{\beta}))$

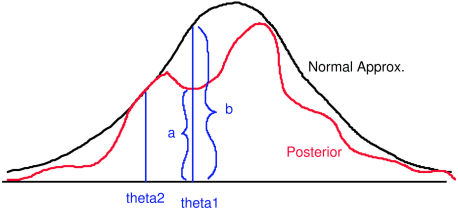
3. EMs: EM for maximization and then simulation (as usual) from asymptotic normal posterior
    (a) EMs adds estimation uncertainty to EM imputations by drawing $\tilde{\beta}$ from its asymptotic normal distribution, $N(\hat{\beta}, \hat{V}(\hat{\beta}))$
    (b) The central limit theorem guarantees that this works as $n \to \infty$, but for real sample sizes it may be inadequate.

3. EMs: EM for maximization and then simulation (as usual) from asymptotic normal posterior
    (a) EMs adds estimation uncertainty to EM imputations by drawing $\tilde{\beta}$ from its asymptotic normal distribution, $N(\hat{\beta}, \hat{V}(\hat{\beta}))$
    (b) The central limit theorem guarantees that this works as $n \to \infty$, but for real sample sizes it may be inadequate.
4. EMis: EM with simulation via importance resampling (probabilistic rejection sampling to draw from the posterior)

3. EMs: EM for maximization and then simulation (as usual) from asymptotic normal posterior
    (a) EMs adds estimation uncertainty to EM imputations by drawing $\tilde{\beta}$ from its asymptotic normal distribution, $N(\hat{\beta}, \hat{V}(\hat{\beta}))$
    (b) The central limit theorem guarantees that this works as $n \to \infty$, but for real sample sizes it may be inadequate.
4. EMis: EM with simulation via importance resampling (probabilistic rejection sampling to draw from the posterior)



3. EMs: EM for maximization and then simulation (as usual) from asymptotic normal posterior
   (a) EMs adds estimation uncertainty to EM imputations by drawing $\tilde{\beta}$ from its asymptotic normal distribution, $N(\hat{\beta}, \hat{V}(\hat{\beta}))$
   (b) The central limit theorem guarantees that this works as $n \to \infty$, but for real sample sizes it may be inadequate.
4. 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.

5. IP: Imputation-posterior

5. IP: Imputation-posterior
   (a) A stochastic version of EM

5. IP: Imputation-posterior
    (a) A stochastic version of EM
    (b) The algorithm. For $\theta = \{\mu, \Sigma\}$,

5. IP: Imputation-posterior
   (a) A stochastic version of EM
   (b) The algorithm. For $\theta = \{\mu, \Sigma\}$,
      i. I-Step: draw D_{mis} from $P(D_{mis}|D_{obs}, \tilde{\theta})$ (i.e., $\tilde{y} = x\tilde{\beta} + \tilde{\epsilon}$) given current draw of $\tilde{\theta}$.

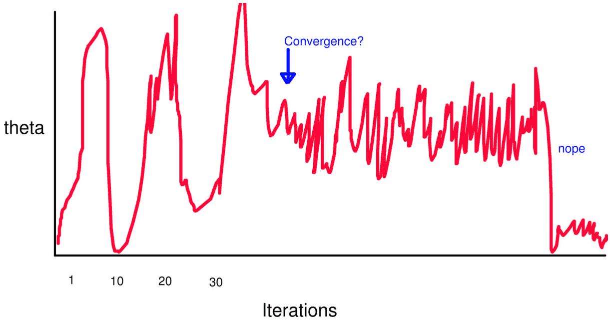
5. IP: Imputation-posterior
   (a) A stochastic version of EM
   (b) The algorithm. For $\theta = \{\mu, \Sigma\}$,
      i. I-Step: draw D_{mis} from $P(D_{mis}|D_{obs}, \tilde{\theta})$ (i.e., $\tilde{y} = x\tilde{\beta} + \tilde{\epsilon}$) given current draw of $\tilde{\theta}$.
      ii. P-Step: draw θ from $P(\theta|D_{obs}, \tilde{D}_{mis})$, given current imputation $\tilde{D}_{mis}$

5. IP: Imputation-posterior
   (a) A stochastic version of EM
   (b) The algorithm. For $\theta = \{\mu, \Sigma\}$,
      i. I-Step: draw D_{mis} from $P(D_{mis}|D_{obs}, \tilde{\theta})$ (i.e., $\tilde{y} = x\tilde{\beta} + \tilde{\epsilon}$) given current draw of $\tilde{\theta}$.
      ii. P-Step: draw θ from $P(\theta|D_{obs}, \tilde{D}_{mis})$, given current imputation $\tilde{D}_{mis}$
   (c) Example of MCMC (Markov-Chain Monte Carlo) methods, one of the most important developments in statistics in the 1990s

5. IP: Imputation-posterior
   (a) A stochastic version of EM
   (b) The algorithm. For $\theta = \{\mu, \Sigma\}$,
      i. I-Step: draw D_{mis} from $P(D_{mis}|D_{obs}, \tilde{\theta})$ (i.e., $\tilde{y} = x\tilde{\beta} + \tilde{\epsilon}$) given current draw of $\tilde{\theta}$.
      ii. P-Step: draw θ from $P(\theta|D_{obs}, \tilde{D}_{mis})$, given current imputation $\tilde{D}_{mis}$
   (c) Example of MCMC (Markov-Chain Monte Carlo) methods, one of the most important developments in statistics in the 1990s
   (d) MCMC enabled statisticians to do things they never previously dreamed possible, but it requires considerable expertise to use and so didn't help others do these things. (Few MCMC routines have appeared in canned packages.)

5. IP: Imputation-posterior
    (a) A stochastic version of EM
    (b) The algorithm. For $\theta = \{\mu, \Sigma\}$,
        i. I-Step: draw D_{mis} from $P(D_{mis}|D_{obs}, \tilde{\theta})$ (i.e., $\tilde{y} = x\tilde{\beta} + \tilde{\epsilon}$) given current draw of $\tilde{\theta}$.
        ii. P-Step: draw θ from $P(\theta|D_{obs}, \tilde{D}_{mis})$, given current imputation $\tilde{D}_{mis}$
    (c) Example of MCMC (Markov-Chain Monte Carlo) methods, one of the most important developments in statistics in the 1990s
    (d) MCMC enabled statisticians to do things they never previously dreamed possible, but it requires considerable expertise to use and so didn't help others do these things. (Few MCMC routines have appeared in canned packages.)
    (e) Hard to know when its over. Convergence is asymptotic in iterations. Plot traces are hard to interpret, and the worst-converging parameter controls the system.

MCMC Convergence Issues



One more algorithm: EMB: EM With Bootstrap

One more algorithm: EMB: EM With Bootstrap

- Randomly draw n obs (with replacement) from the data

One more algorithm: EMB: EM With Bootstrap

- Randomly draw n obs (with replacement) from the data
- Use EM to estimate β and Σ in each (for estimation uncertainty)

One more algorithm: EMB: EM With Bootstrap

- Randomly draw n obs (with replacement) from the data
- Use EM to estimate β and Σ in each (for estimation uncertainty)
- Impute D_{mis} from each from the model (for fundamental uncertainty)

One more algorithm: EMB: EM With Bootstrap

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

One more algorithm: EMB: EM With Bootstrap

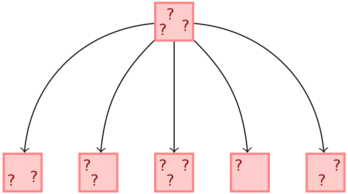
- Randomly draw n obs (with replacement) from the data
- Use EM to estimate β and Σ in each (for estimation uncertainty)
- Impute D_{mis} from each from the model (for fundamental uncertainty)
- Lightning fast; works with very large data sets
- Basis for Amelia II

Multiple Imputation: Amelia Style

Multiple Imputation: Amelia Style

incomplete data

Multiple Imputation: Amelia Style

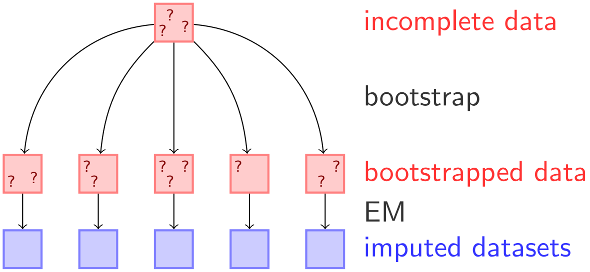


incomplete data

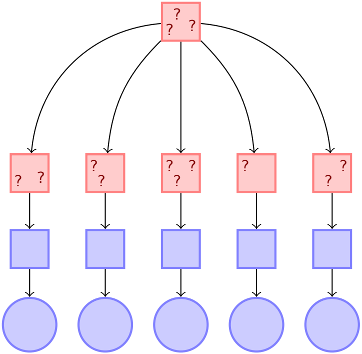
bootstrap

bootstrapped data

Multiple Imputation: Amelia Style



Multiple Imputation: Amelia Style



incomplete data

bootstrap

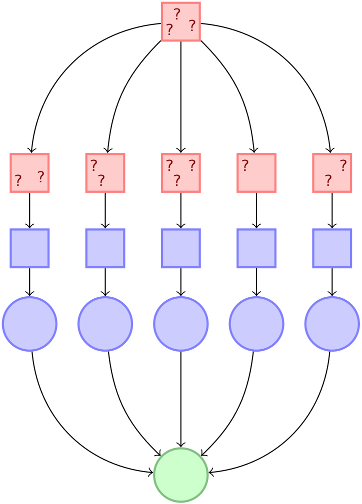
bootstrapped data

EM

imputed datasets

analysis

Multiple Imputation: Amelia Style



incomplete data

bootstrap

bootstrapped data

EM

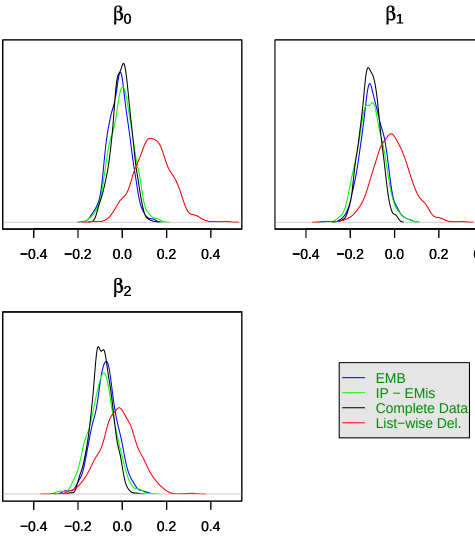
imputed datasets

analysis

combination

final results

Comparisons of Posterior Density Approximations



What Can Go Wrong and What to Do

What Can Go Wrong and What to Do

- 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

- Inference is learning about facts we don't have with facts we have; we *assume* the 2 are related!
- Imputation and analysis are estimated separately $\rightsquigarrow$ robustness because imputation affects only missing observations. High missingness reduces the property.

What Can Go Wrong and What to Do

- Inference is learning about facts we don't have with facts we have; we *assume* the 2 are related!
- Imputation and analysis are estimated separately $\rightsquigarrow$ robustness because imputation affects only missing observations. High missingness reduces the property.
- 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

- Inference is learning about facts we don't have with facts we have; we *assume* the 2 are related!
- Imputation and analysis are estimated separately $\rightsquigarrow$ robustness because imputation affects only missing observations. High missingness reduces the property.
- 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)
- Fit imputation model distributional assumptions by transformation to unbounded scales: $\sqrt{\text{counts}}$, $\ln(p/(1-p))$, $\ln(\text{money})$, etc.

What Can Go Wrong and What to Do

- Inference is learning about facts we don't have with facts we have; we *assume* the 2 are related!
- Imputation and analysis are estimated separately $\rightsquigarrow$ robustness because imputation affects only missing observations. High missingness reduces the property.
- 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)
- Fit imputation model distributional assumptions by transformation to unbounded scales: $\sqrt{\text{counts}}$, $\ln(p/(1-p))$, $\ln(\text{money})$, 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)

- Represent severe nonlinear relationships in the imputation model with transformations or added quadratic terms.

What Can Go Wrong and What to Do (continued)

- Represent severe nonlinear relationships in the imputation model with transformations or added quadratic terms.
- If imputation model has as much information as the 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)

- Represent severe nonlinear relationships in the imputation model with transformations or added quadratic terms.
- If imputation model has as much information as the analysis model, but the specification (such as the functional form) differs, CIs are conservative (e.g., $\geq 95\%$ CIs)
- When imputation model includes more information than analysis model, it can be more efficient than the "optimal" application-specific model (known as "super-efficiency")

What Can Go Wrong and What to Do (continued)

- Represent severe nonlinear relationships in the imputation model with transformations or added quadratic terms.
- If imputation model has as much information as the analysis model, but the specification (such as the functional form) differs, CIs are conservative (e.g., $\geq 95\%$ CIs)
- When imputation model includes more information than analysis model, it can be more efficient than the "optimal" application-specific model (known as "super-efficiency")
- Bad intuitions

What Can Go Wrong and What to Do (continued)

- Represent severe nonlinear relationships in the imputation model with transformations or added quadratic terms.
- If imputation model has as much information as the analysis model, but the specification (such as the functional form) differs, CIs are conservative (e.g., $\geq$ 95% CIs)
- When imputation model includes more information than analysis model, it can be more efficient than the "optimal" application-specific model (known as "super-efficiency")
- Bad intuitions
  - If X is randomly imputed why no attenuation (the usual consequence of random measurement error in an explanatory variable)?

What Can Go Wrong and What to Do (continued)

- Represent severe nonlinear relationships in the imputation model with transformations or added quadratic terms.
- If imputation model has as much information as the analysis model, but the specification (such as the functional form) differs, CIs are conservative (e.g., $\geq 95\%$ CIs)
- When imputation model includes more information than analysis model, it can be more efficient than the "optimal" application-specific model (known as "super-efficiency")
- Bad intuitions
  - If X is randomly imputed why no attenuation (the usual consequence of random measurement error in an explanatory variable)?
  - If X is imputed with information from Y, why no endogeneity?

What Can Go Wrong and What to Do (continued)

- Represent severe nonlinear relationships in the imputation model with transformations or added quadratic terms.
- If imputation model has as much information as the analysis model, but the specification (such as the functional form) differs, CIs are conservative (e.g., $\geq$ 95% CIs)
- When imputation model includes more information than analysis model, it can be more efficient than the "optimal" application-specific model (known as "super-efficiency")
- Bad intuitions
  - If X is randomly imputed why no attenuation (the usual consequence of random measurement error in an explanatory variable)?
  - If X is imputed with information from Y, why no endogeneity?
  - Answer to both: the draws are from the joint posterior and put back into the data. Nothing is being changed.

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:

1. 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:

1. 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.).
2. 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:

1. 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.).
2. NI missingness in X and no external variables are available that could be used in an imputation stage to fix the problem.
3. 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:

1. 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.).
2. NI missingness in X and no external variables are available that could be used in an imputation stage to fix the problem.
3. Missingness in X is not a function of Y
4. The n left after listwise deletion is so large that the efficiency loss does not counter balance the biases induced by the other conditions.

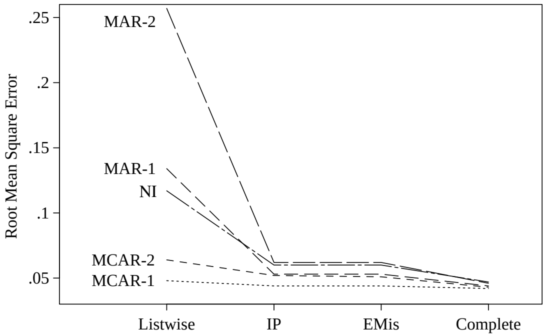
The Best Case for Listwise Deletion

Listwise deletion is better than MI when all 4 hold:

1. 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.).
2. NI missingness in X and no external variables are available that could be used in an imputation stage to fix the problem.
3. Missingness in X is not a function of Y
4. The n left after listwise deletion is so large that the efficiency loss does not counter balance the biases induced by the other conditions.

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

Root Mean Square Error Comparisons



Each point is RMSE averaged over two regression coefficients in each of 100 simulated data sets. (IP and EMis have the same RMSE, which is lower than listwise deletion and higher than the complete data; its the same for EMB.)

Detailed Example: Support for Perot

Detailed Example: Support for Perot

1. Research question: were voters who did not share in the economic recovery more likely to support Perot in the 1996 presidential election?

Detailed Example: Support for Perot

1. Research question: were voters who did not share in the economic recovery more likely to support Perot in the 1996 presidential election?
2. Analysis model: linear regression

Detailed Example: Support for Perot

1. Research question: were voters who did not share in the economic recovery more likely to support Perot in the 1996 presidential election?
2. Analysis model: linear regression
3. Data: 1996 National Election Survey (n=1714)

Detailed Example: Support for Perot

1. Research question: were voters who did not share in the economic recovery more likely to support Perot in the 1996 presidential election?
2. Analysis model: linear regression
3. Data: 1996 National Election Survey (n=1714)
4. Dependent variable: Perot Feeling Thermometer

Detailed Example: Support for Perot

1. Research question: were voters who did not share in the economic recovery more likely to support Perot in the 1996 presidential election?
2. Analysis model: linear regression
3. Data: 1996 National Election Survey (n=1714)
4. Dependent variable: Perot Feeling Thermometer
5. Key explanatory variables: retrospective and propsective evaluations of national economic performance and personal financial circumstances

Detailed Example: Support for Perot

1. Research question: were voters who did not share in the economic recovery more likely to support Perot in the 1996 presidential election?
2. Analysis model: linear regression
3. Data: 1996 National Election Survey (n=1714)
4. Dependent variable: Perot Feeling Thermometer
5. Key explanatory variables: retrospective and propsective evaluations of national economic performance and personal financial circumstances
6. Control variables: age, education, family income, race, gender, union membership, ideology

Detailed Example: Support for Perot

1. Research question: were voters who did not share in the economic recovery more likely to support Perot in the 1996 presidential election?
2. Analysis model: linear regression
3. Data: 1996 National Election Survey (n=1714)
4. Dependent variable: Perot Feeling Thermometer
5. Key explanatory variables: retrospective and propsective evaluations of national economic performance and personal financial circumstances
6. Control variables: age, education, family income, race, gender, union membership, ideology
7. Extra variables included in the 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.

Detailed Example: Support for Perot

1. Research question: were voters who did not share in the economic recovery more likely to support Perot in the 1996 presidential election?
2. Analysis model: linear regression
3. Data: 1996 National Election Survey (n=1714)
4. Dependent variable: Perot Feeling Thermometer
5. Key explanatory variables: retrospective and propsective evaluations of national economic performance and personal financial circumstances
6. Control variables: age, education, family income, race, gender, union membership, ideology
7. Extra variables included in the 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.
8. Include nonlinear terms: age^2

9. Transform variables to more closely approximate distributional assumptions: logged number of organizations participating in.

9. Transform variables to more closely approximate distributional assumptions: logged number of organizations participating in.
10. Run Amelia to generate 5 imputed data sets.

9. Transform variables to more closely approximate distributional assumptions: logged number of organizations participating in.
10. Run Amelia to generate 5 imputed data sets.
11. Key substantive result is the coefficient on retrospective economic evaluations (ranges from 1 to 5):

9. Transform variables to more closely approximate distributional assumptions: logged number of organizations participating in.
10. Run Amelia to generate 5 imputed data sets.
11. Key substantive result is the coefficient on retrospective economic evaluations (ranges from 1 to 5):

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

9. Transform variables to more closely approximate distributional assumptions: logged number of organizations participating in.
10. Run Amelia to generate 5 imputed data sets.
11. Key substantive result is the coefficient on retrospective economic evaluations (ranges from 1 to 5):

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

so $(5-1) \times 1.65 = 6.6$, which is also a percentage of the range of Y.

9. Transform variables to more closely approximate distributional assumptions: logged number of organizations participating in.
10. Run Amelia to generate 5 imputed data sets.
11. Key substantive result is the coefficient on retrospective economic evaluations (ranges from 1 to 5):

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

so $(5-1)\times 1.65 = 6.6$, which is also a percentage of the range of Y.

(a) MI estimator is more efficient, with a smaller SE

9. Transform variables to more closely approximate distributional assumptions: logged number of organizations participating in.
10. Run Amelia to generate 5 imputed data sets.
11. Key substantive result is the coefficient on retrospective economic evaluations (ranges from 1 to 5):

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

so $(5-1)\times 1.65 = 6.6$, which is also a percentage of the range of Y.

(a) MI estimator is more efficient, with a smaller SE
(b) The MI estimator is 4 times larger

9. Transform variables to more closely approximate distributional assumptions: logged number of organizations participating in.
10. Run Amelia to generate 5 imputed data sets.
11. Key substantive result is the coefficient on retrospective economic evaluations (ranges from 1 to 5):

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

so $(5-1) \times 1.65 = 6.6$, which is also a percentage of the range of Y.

(a) MI estimator is more efficient, with a smaller SE
(b) The MI estimator is 4 times larger
(c) Based on listwise deletion, there is no evidence that perception of poor economic performance is related to support for Perot

9. Transform variables to more closely approximate distributional assumptions: logged number of organizations participating in.
10. Run Amelia to generate 5 imputed data sets.
11. Key substantive result is the coefficient on retrospective economic evaluations (ranges from 1 to 5):

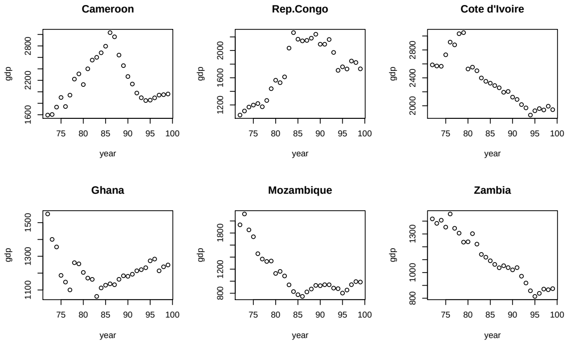
Listwise deletion	.43
	(.90)
Multiple imputation	1.65
	(.72)

so $(5-1) \times 1.65 = 6.6$, which is also a percentage of the range of Y.

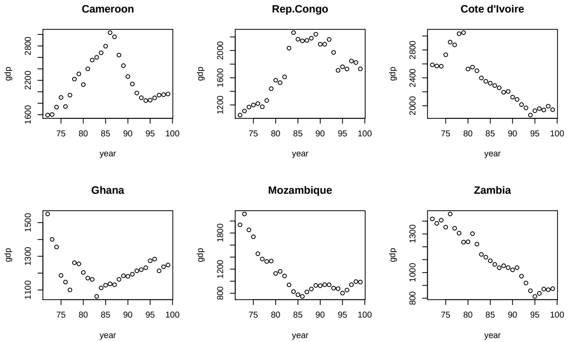
(a) MI estimator is more efficient, with a smaller SE
(b) The MI estimator is 4 times larger
(c) Based on listwise deletion, there is no evidence that perception of poor economic performance is related to support for Perot
(d) Based on the MI estimator, R's with negative retrospective economic evaluations are more likely to have favorable views of Perot.

MI in Time Series Cross-Section Data

MI in Time Series Cross-Section Data

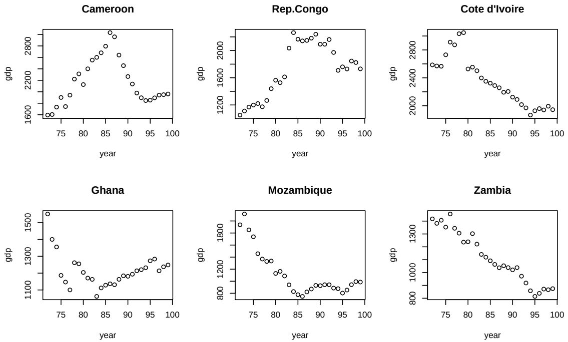


MI in Time Series Cross-Section Data



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

MI in Time Series Cross-Section Data

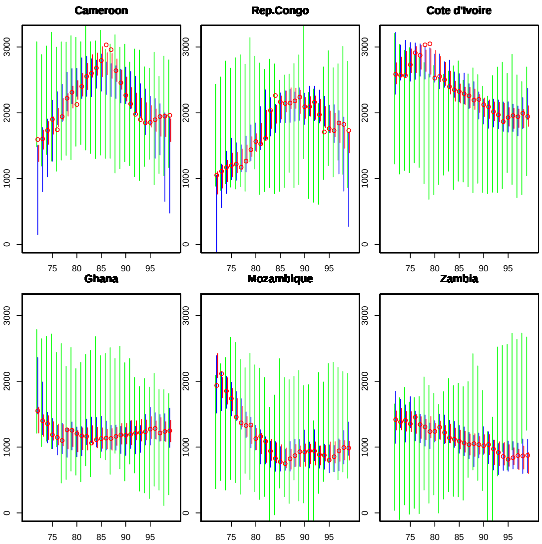


Include: (1) fixed effects, (2) time trends, and (3) priors for cells

Read: James Honaker and Gary King, "What to do About Missing Values in Time Series Cross-Section Data,"

http://gking.harvard.edu/files/abs/pr-abs.shtml

Imputation one Observation at a time



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

Priors on Cell Values

Priors on Cell Values

- Recall: $p(\theta|y) = p(\theta)\prod_{i=1}^{n} L_i(\theta|y)$

Priors on Cell Values

- Recall: $p(\theta|y) = p(\theta)\prod_{i=1}^{n} L_i(\theta|y)$
- take logs: $\ln p(\theta|y) = \ln[p(\theta)] + \sum_{i=1}^{n} \ln L_i(\theta|y)$

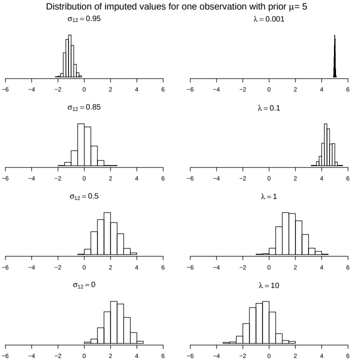
Priors on Cell Values

- Recall: $p(\theta|y) = p(\theta)\prod_{i=1}^{n} L_i(\theta|y)$
- take logs: $\ln p(\theta|y) = \ln[p(\theta)] + \sum_{i=1}^{n} \ln L_i(\theta|y)$
- $\Longrightarrow$ Suppose prior is of the same form, $p(\theta|y) = L_i(\theta|y)$; then its just another observation: $\ln p(\theta|y) = \sum_{i=1}^{n+1} \ln L_i(\theta|y)$

Priors on Cell Values

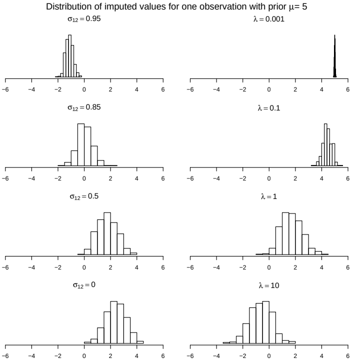
- Recall: $p(\theta|y) = p(\theta)\prod_{i=1}^{n} L_i(\theta|y)$
- take logs: $\ln p(\theta|y) = \ln[p(\theta)] + \sum_{i=1}^{n} \ln L_i(\theta|y)$
- $\Longrightarrow$ Suppose prior is of the same form, $p(\theta|y) = L_i(\theta|y)$; then its just another observation: $\ln p(\theta|y) = \sum_{i=1}^{n+1} \ln L_i(\theta|y)$
- Honaker and King show how to modify these "data augmentation priors" to put priors on missing values rather than on μ and σ (or β).

Posterior imputation: mean=0, prior mean=5



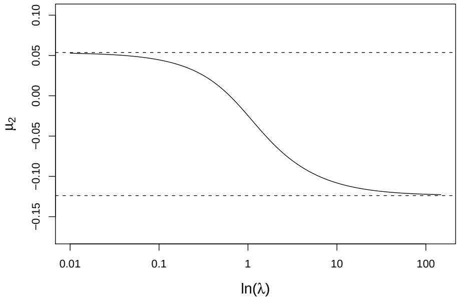
Left column: holds prior $N(5, \lambda)$ constant ($\lambda = 1$) and changes predictive strength (the covariance, σ_{12}).

Posterior imputation: mean=0, prior mean=5



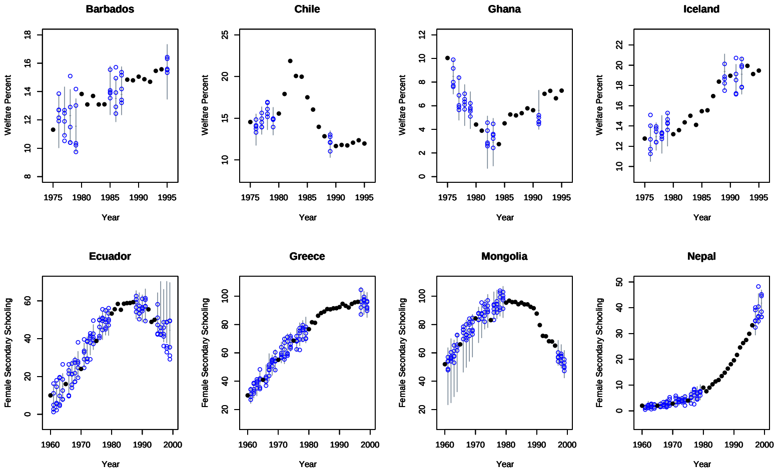
Left column: holds prior $N(5, \lambda)$ constant ($\lambda = 1$) and changes predictive strength (the covariance, σ_{12}).
Right column: holds predictive strength of data constant (at $\sigma_{12} = 0.5$) and changes the strength of the prior (λ).

Model Parameters Respond to Prior on a Cell Value



Prior: $p(x_{12}) = N(5, \lambda)$. The parameter approaches the theoretical limits (dashed lines), upper bound is what is generated when the missing value is filled in with the expectation; lower bound is the parameter when the model is estimated without priors. The overall movement is small.

Replication of Baum and Lake; Imputation Model Fit



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

	Listwise Deletion	Multiple Imputation
Life Expectancy		
Rich Democracies	−.072	.233
	(.179)	(.037)
Poor Democracies	−.082	.120
	(.040)	(.099)
N	1789	5627
Secondary Education		
Rich Democracies	.948	.948
	(.002)	(.019)
Poor Democracies	.373	.393
	(.094)	(.081)
N	1966	5627

Replication of Baum and Lake; the effect of being a democracy on life expectancy and on the percentage enrolled in secondary education (with p-values in parentheses).