Lecture 14: Missing Data

Naijia Liu

March 19 2024

Logistics

- Midterm solution set will be released on Thursday before lecture time.
- Problem Set II will be available on Thursday, and due next Thursday at **midnight**.

Logistics

- Midterm solution set will be released on Thursday before lecture time.
- Problem Set II will be available on Thursday, and due next Thursday at **midnight**.
- **By April 4th,** mandatory OH with Naijia, Jeremiah or James, per group.

  We'd like to hear about your idea and dataset.
- If you are still looking for data:

  Run `data()` in `R-studio`.

  These are built-in and cleaned datasets!

Why is our data missing?

- What is your household income in the year of 2022?

  Extremely rich people may refuse to answer.

Why is our data missing?

- What is your household income in the year of 2022?

  Extremely rich people may refuse to answer.

- What is your lowest score of a college class?

  A failing grade does not look good.

Why is our data missing?

- What is your household income in the year of 2022?

  Extremely rich people may refuse to answer.

- What is your lowest score of a college class?

  A failing grade does not look good.

- Have you committed a crime before?

  There will be consequence if yes.

Why is our data missing?

- What is your household income in the year of 2022?

  Extremely rich people may refuse to answer.
- What is your lowest score of a college class?

  A failing grade does not look good.
- Have you committed a crime before?

  There will be consequence if yes.
- What was the CO2 amount of every country in 1990?

  Governments did not document the data / chose not to report (countries want to hide their CO2 omission).

Why is missing data a problem?

- What is your household income in the year of 2022?

  We lose the richest group in our analysis.

Why is missing data a problem?

- What is your household income in the year of 2022?

  We lose the richest group in our analysis.

- What is your lowest score of a college class?

  Grade distribution would be biased towards higher grades.

Why is missing data a problem?

- What is your household income in the year of 2022?

  We lose the richest group in our analysis.

- What is your lowest score of a college class?

  Grade distribution would be biased towards higher grades.

- Have you committed a crime before?

  We want to be able to catch criminals!

Why is missing data a problem?

- What is your household income in the year of 2022?

  We lose the richest group in our analysis.

- What is your lowest score of a college class?

  Grade distribution would be biased towards higher grades.

- Have you committed a crime before?

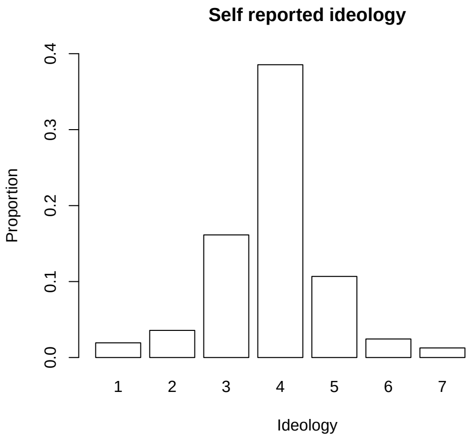
  We want to be able to catch criminals!

- What was the CO_2 amount of every country in 1990?

  We want to be able to study all types countries.

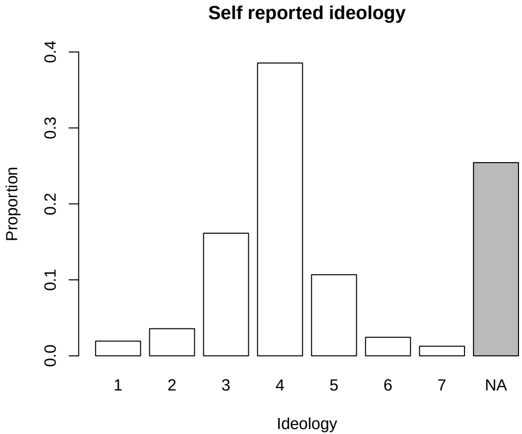
Why is missing data a problem?

- Survey questions to Chinese respondents: what is your ideology?
  - Anti China Communist Party respondents self censor, fearing consequences.



Why is missing data a problem?

- Survey questions to Chinese respondents: what is your ideology?
  - Anti China Communist Party respondents self censor, fearing consequences.
  - What if we completely miss a type of respondents in our analysis?



How do we think of missing data?

- Missing completely at random

How do we think of missing data?

- Missing completely at random
  - Imagine spilling coffee onto the data sheet.

How do we think of missing data?

- Missing completely at random
  - Imagine spilling coffee onto the data sheet.
  - Randomly choose Chinese respondents to refuse to answer the ideology question.

How do we think of missing data?

- Missing completely at random
  - Imagine spilling coffee onto the data sheet.
  - Randomly choose Chinese respondents to refuse to answer the ideology question.
  - Listwise deletion can deal with MCAR.

    i.e get rid of those who refused to answer the ideology question.

Missing Completely at Random

- MCAR is not plausible in reality.

  Even with spilling coffee, variables / observations closer to coffee mugs are more likely to go missing.

Missing Completely at Random

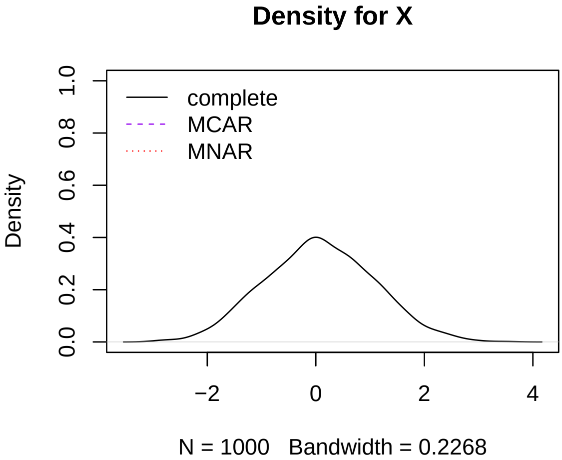
- MCAR is not plausible in reality.

  Even with spilling coffee, variables / observations closer to coffee mugs are more likely to go missing.

  People with extreme ideology may feel insecure to reveal it.

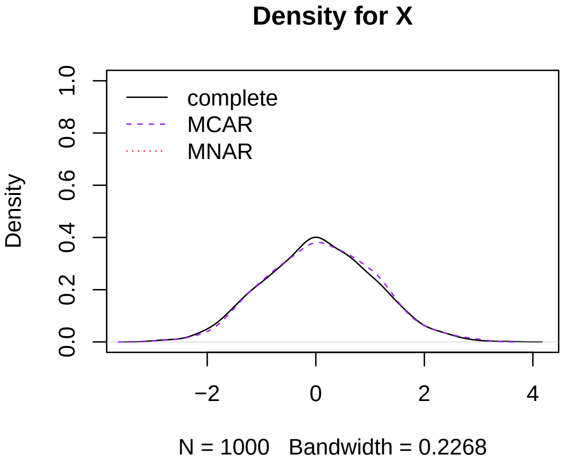
Missing Completely at Random

- If MCAR is true, we can delete observations as if we only get a smaller sample of the same population.



Missing Completely at Random

- If MCAR is true, we can delete observations as if we only get a smaller sample of the same population.



How do we think of missing data?

- Missing at random
  - Conditioning on observables, missing values and observed values are similar in general.

How do we think of missing data?

- Missing at random
  - Conditioning on observables, missing values and observed values are similar in general.
  - **Conditioning on all other variables in the dataset (such as age, gender, education), missing ideology answers are similar to observed responses, on average.**

Missing at Random

- Missing at random is more plausible than missing completely at random.
- We allow missing values to be different from observed values. The differences go away after taking into consideration of the observed variables.
- This indicates that we can utilize observed info to **impute** missing values.

Multiple Imputation and Missing at Random

- Say we start with missing value in ideology variable only.

Multiple Imputation and Missing at Random

- Say we start with missing value in ideology variable only.
  - Observed: age, gender, education, ideology (only partially)

Multiple Imputation and Missing at Random

- Say we start with missing value in ideology variable only.
  - Observed: age, gender, education, ideology (only partially)
  - Missing: ideology (only partially)

Multiple Imputation and Missing at Random

- Say we start with missing value in ideology variable only.
  - Observed: age, gender, education, ideology (only partially)
  - Missing: ideology (only partially)
- We train a linear regression model using complete cases:

$$\text{Ideology} = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{gender} + \beta_3 \cdot \text{edu} + \epsilon$$

Multiple Imputation and Missing at Random

- Say we start with missing value in ideology variable only.
  - Observed: age, gender, education, ideology (only partially)
  - Missing: ideology (only partially)
- We train a linear regression model using complete cases:

$$\text{Ideology} = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{gender} + \beta_3 \cdot \text{edu} + \epsilon$$

- We **impute** / predict missing ideology answers using this linear model.

Multiple Imputation and Missing at Random

- Say we start with missing value in ideology variable only.
  - Observed: age, gender, education, ideology (only partially)
  - Missing: ideology (only partially)
- We train a linear regression model using complete cases:

$$\text{Ideology} = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{gender} + \beta_3 \cdot \text{edu} + \epsilon$$

- We **impute** / predict missing ideology answers using this linear model.
- Data is now complete.

Multiple imputation

1. A simple imputation, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as "place holders."

Multiple imputation

1. A simple imputation, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as "place holders."
2. The "place holder" mean imputations for one variable ("var") are set back to missing.

Multiple imputation

1. A simple imputation, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as "place holders."
2. The "place holder" mean imputations for one variable ("var") are set back to missing.
3. The observed values from the variable "var" in Step 2 are regressed on the other variables in the imputation model. In other words, "var" is the dependent variable in a regression model and all the other variables are independent variables in the regression model.

Multiple imputation

1. A simple imputation, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as "place holders."
2. The "place holder" mean imputations for one variable ("var") are set back to missing.
3. The observed values from the variable "var" in Step 2 are regressed on the other variables in the imputation model. In other words, "var" is the dependent variable in a regression model and all the other variables are independent variables in the regression model.
4. The missing values for "var" are then replaced with predictions (imputations) from the regression model.

Multiple imputation

1. A simple imputation, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as "place holders."
2. The "place holder" mean imputations for one variable ("var") are set back to missing.
3. The observed values from the variable "var" in Step 2 are regressed on the other variables in the imputation model. In other words, "var" is the dependent variable in a regression model and all the other variables are independent variables in the regression model.
4. The missing values for "var" are then replaced with predictions (imputations) from the regression model.
5. Steps 2–4 are then repeated for each variable that has missing data.

Multiple imputation

1. A simple imputation, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as "place holders."
2. The "place holder" mean imputations for one variable ("var") are set back to missing.
3. The observed values from the variable "var" in Step 2 are regressed on the other variables in the imputation model. In other words, "var" is the dependent variable in a regression model and all the other variables are independent variables in the regression model.
4. The missing values for "var" are then replaced with predictions (imputations) from the regression model.
5. Steps 2–4 are then repeated for each variable that has missing data.
6. Steps 2–4 are repeated for a number of cycles, with the imputations being updated at each cycle.

Multiple imputation

- Assumptions: Missing at Random.

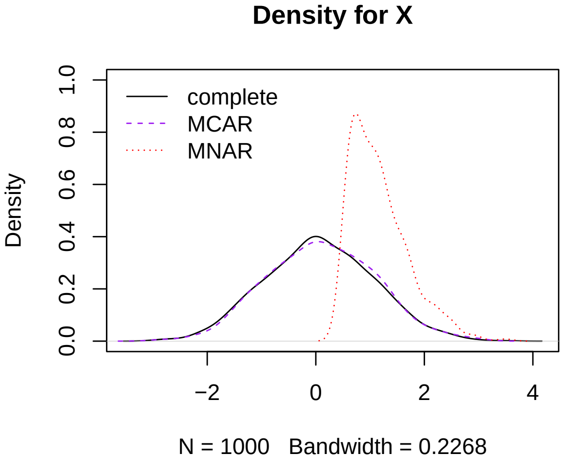
  We utilize other observed variables to impute.

- Usually produce different results with different starting point.

  One solution is to take average among the multiply imputed datasets.

How do we think of missing data?

- Missing **NOT** at random
    - Systematic selection leads to missing values.



Missing NOT at Random

- Missing not at random is very possible in social science datasets.

Missing NOT at Random

- Missing not at random is very possible in social science datasets.
- Sensitive survey questions.

Missing NOT at Random

- Missing not at random is very possible in social science datasets.
- Sensitive survey questions.
- Selective reporting by government / institution.

Missing NOT at Random

- Missing not at random is very possible in social science datasets.
- Sensitive survey questions.
- Selective reporting by government / institution.
- Listwise deletion and multiple imputation cannot solve MNAR.

  Because we need more information about the systematic selection. These info are not in the observed variables.

Missing NOT at Random

- Transparency, Protest and Democratic Stability (Hollyer et al, BJPS, 2018)
- Measure transparency through missing data in country reports.
- Transparency is associated with a reduction in both the probability of democratic collapse and of the irregular removal of democratic leaders. Transparency stabilizes democratic rule.

Hollyer et al, 2018

- The availability of information on aggregate policy outcomes.
- Authors define transparency as a latent predictor of the reporting/non-reporting of data to the World Bank's World Development Indicators (WDI) data series.

Variable	Mean	Stand. Dev.	Min.	Max.
Transparency	2.50	2.19	−1.37	9.98
Growth (pct. GDP)	1.81	4.24	−26.2	31.9
GDP *per capita* (thousands 2005 PPP USD)	12.8	10.4	0.37	46.7
Ec. Openness (pct. GDP)	64.6	34.5	10.3	222
Parliamentary	0.42	0.49	0	1
Mixed System	0.18	0.38	0	1

Hollyer et al, 2018

- Listwise delete non-reported data ?

Hollyer et al, 2018

- Listwise delete non-reported data ?

  Assuming complete randomness in non-reporting.

Hollyer et al, 2018

- Listwise delete non-reported data ?

  Assuming complete randomness in non-reporting.
- Multiple impute the non-reported data?

Hollyer et al, 2018

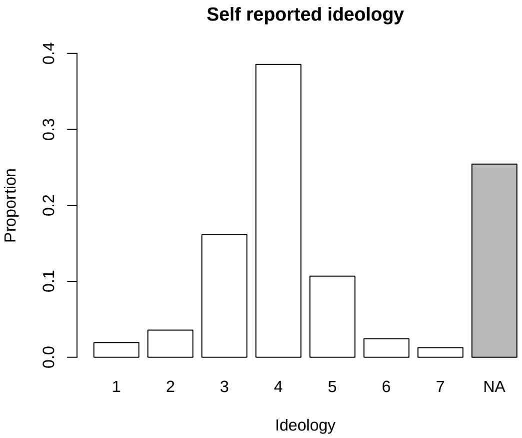
- Listwise delete non-reported data ?

  Assuming complete randomness in non-reporting.

- Multiple impute the non-reported data?

  Assuming randomness in non-reporting, conditioning on GDP, Growth and political system.

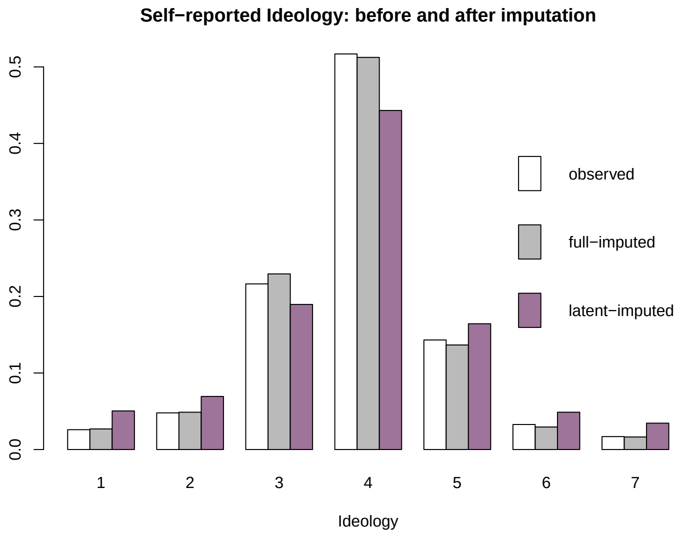
Liu, 2022



Self reported ideology
Proportion
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
6
7
NA
Ideology

Liu, 2022

Chinese respondents with extreme ideology are less likely to report.



Summary

- Missing data is everywhere!

Summary

- Missing data is everywhere!
- Three possible mechanisms:

Summary

- Missing data is everywhere!
- Three possible mechanisms:
  - Missing completely at random

    $\rightsquigarrow$ listwise deletion

Summary

- Missing data is everywhere!
- Three possible mechanisms:
  - Missing completely at random

    $\rightsquigarrow$ listwise deletion
  - Missing at random $\rightsquigarrow$ multiple imputation

Summary

- Missing data is everywhere!
- Three possible mechanisms:
  - Missing completely at random

    $\rightsquigarrow$ listwise deletion
  - Missing at random $\rightsquigarrow$ multiple imputation
  - Missing not at random

    $\rightsquigarrow$ more careful modeling

Summary

- Missing data is everywhere!
- Three possible mechanisms:
  - Missing completely at random

    $\rightsquigarrow$ listwise deletion
  - Missing at random $\rightsquigarrow$ multiple imputation
  - Missing not at random

    $\rightsquigarrow$ more careful modeling
- Dealing with missing values often leads to different study results!

How multiple imputation makes a difference? (Lall, 2017)

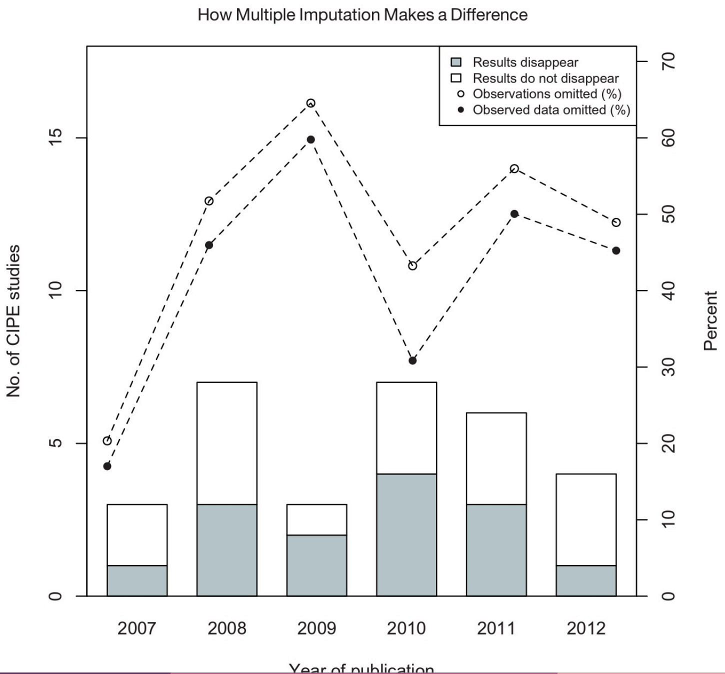
- Large-scale examination of the empirical effects of substituting multiple imputation for listwise deletion in political science.

How multiple imputation makes a difference? (Lall, 2017)

- Large-scale examination of the empirical effects of substituting multiple imputation for listwise deletion in political science.
- Focuses on research in the major subfield of comparative and international political economy (CIPE).

How multiple imputation makes a difference? (Lall, 2017)

- Large-scale examination of the empirical effects of substituting multiple imputation for listwise deletion in political science.
- Focuses on research in the major subfield of comparative and international political economy (CIPE).
- In almost half of the studies, key results “disappear” (by conventional statistical standards) when reanalyzed.


How Multiple Imputation Makes a Difference
Results disappear
Results do not disappear
Observations omitted (%)
Observed data omitted (%)
No. of CIPE studies
0
5
10
15
Percent
0
10
20
30
40
50
60
70
2007
2008
2009
2010
2011
2012
Year of publication