14: More on Linear Models

Naijia Liu

Spring 2025

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.

- What is your lowest score of a college class?

  Grade distribution would be biased towards higher grades.

- Have you committed a crime before?

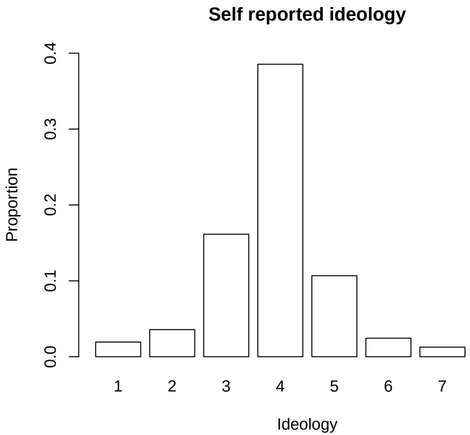
  We want to be able to catch criminals!

- What was the CO2 amount of every country in 1990?

  We want to be able to study all types countries.

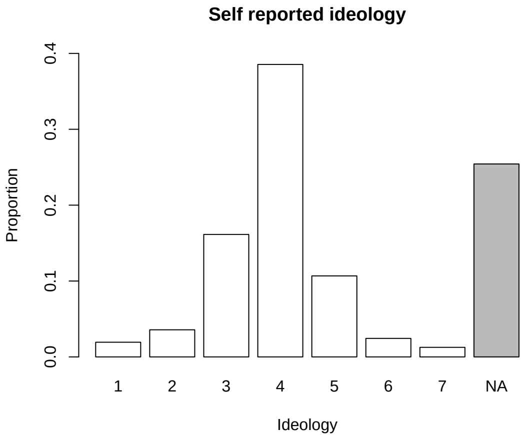
Why is missing data a problem?

- Survey questions to Chinese respondents: what is your ideology?
  - What are the reasons for people not to report?



Why is missing data a problem?

- Survey questions to Chinese respondents: what is your ideology?
  - People with extreme ideology might not feel safe to report.



Roadmap

1. Missing data mechanisms.
2. Missing data on observational studies.
3. Applications and examples.
4. A new estimator for missing confounders in observational studies.

Notation

- X: Complete values
- R: Missingness indicator.
  - $R = 1$ if observed.
  - $R = 0$ if not.
- X_{obs}: Observed values.
- X_{mis}: Missing values.

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.
- Unconditional randomness:

$$X_{\text{mis}} \perp\!\!\!\perp R$$

Missing Completely at Random

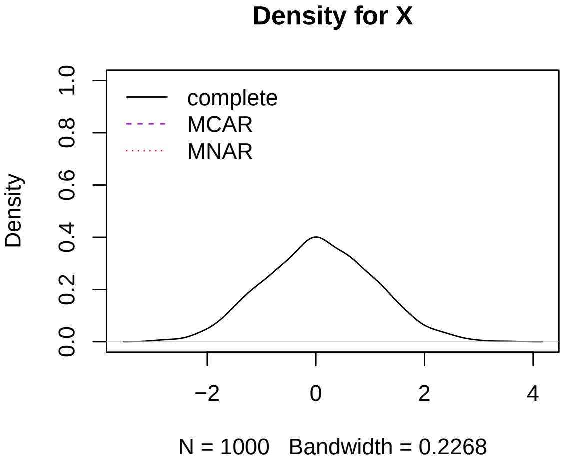
- MCAR is not plausible in reality.

  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.

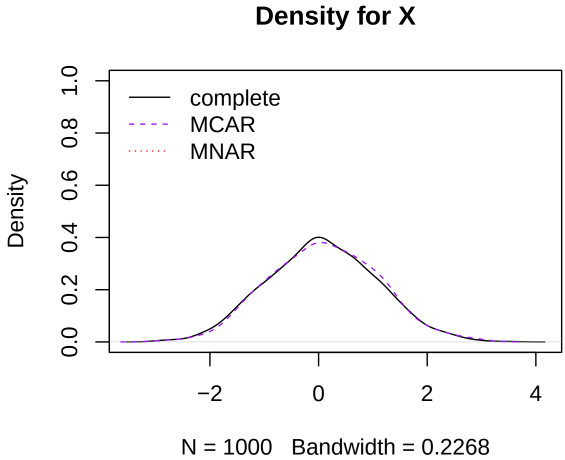
Missing Completely at Random

- 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.
  - **Conditioning on all other variables in the dataset (such as age, gender, education), missing ideology answers are similar to observed responses, on average.**
- Conditional randomness:

$$X_{\mathsf{mis}} \perp\!\!\!\perp R \mid X_{\mathsf{obs}}$$

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.
  - 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."
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.

- `R` pakcage: `Amelia`, `mice` and many more.

Simulation Overview

- Goal: Study the performance of regression estimators under missing data
- Compare three approaches:
  - Oracle (no missingness)
  - Complete case analysis
  - Multiple imputation (MI)

Data Generation Process

- For each simulation ($n = 1000$ observations):
  - Generate $X_1 \sim \mathcal{N}(-4, 0.5)$
  - Generate $X_2 = 0.5X_1 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$
  - Generate $Y = 1 + 2X_1 - 1X_2 + \eta$, where $\eta \sim \mathcal{N}(0, 1)$

Introducing Missingness

- Create missing values in X_1 based on X_2:
- Missingness probability: $\Pr(X_1 \text{ missing}) = \text{logit}^{-1}(X_2 + e)$, where $e \sim \mathcal{N}(1, 1)$
- This creates Missing At Random (MAR) structure

Estimation Procedures

- For each simulated dataset:
  - **Oracle**: Regress Y on X_1 and X_2 using full data
  - **Complete Case**: Regress using only complete observations
  - **Multiple Imputation**: Impute missing X_2 values using MICE (5 imputations) with **pooled regression results**.

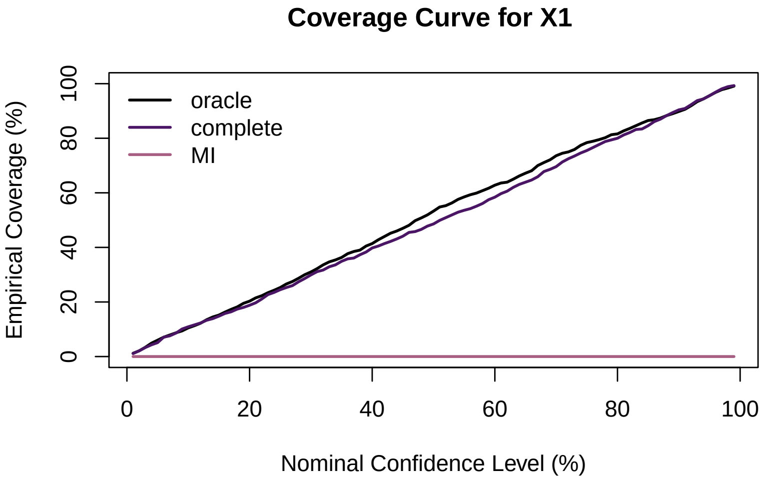
Recorded Results

- For each estimator, we store:
  - Point estimates for β_{X_1} and β_{X_2}
  - Standard errors for β_{X_1} and β_{X_2}
- Total of 1000 simulations

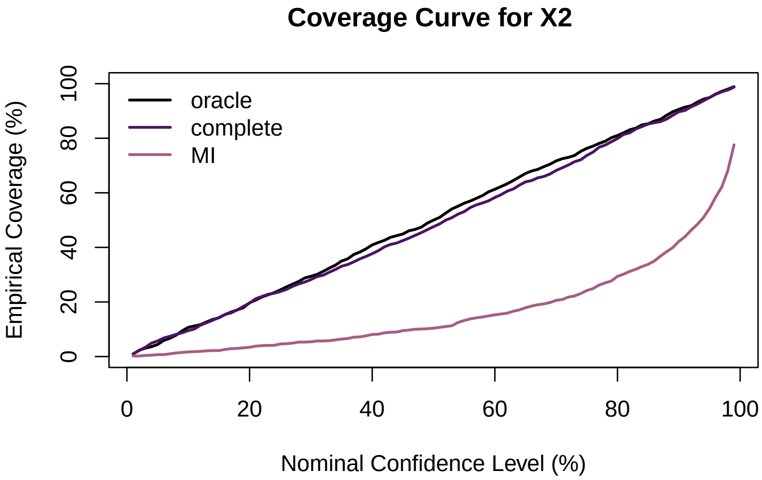
Evaluating Coverage

- For nominal confidence levels from 1% to 99%:
- Construct confidence intervals:
  - $\hat{\beta} \pm t_{\alpha/2,df} \times \widehat{SE}$
- Check whether true parameter falls inside CI
- Calculate empirical coverage rate at each level

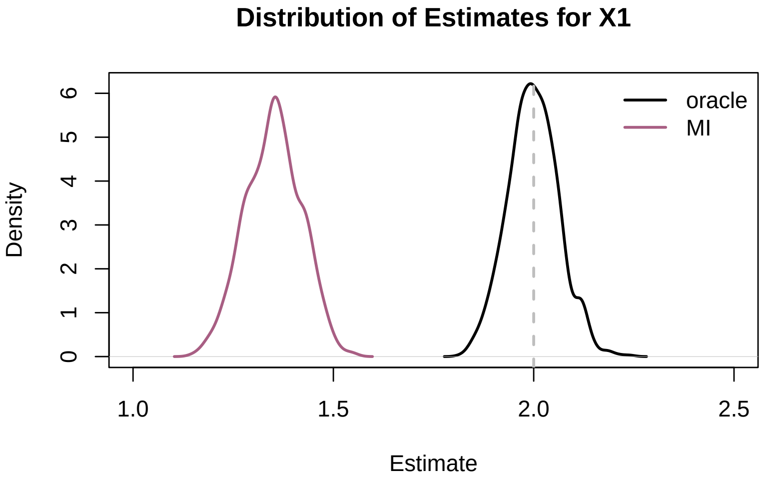
Results: MI under covers!



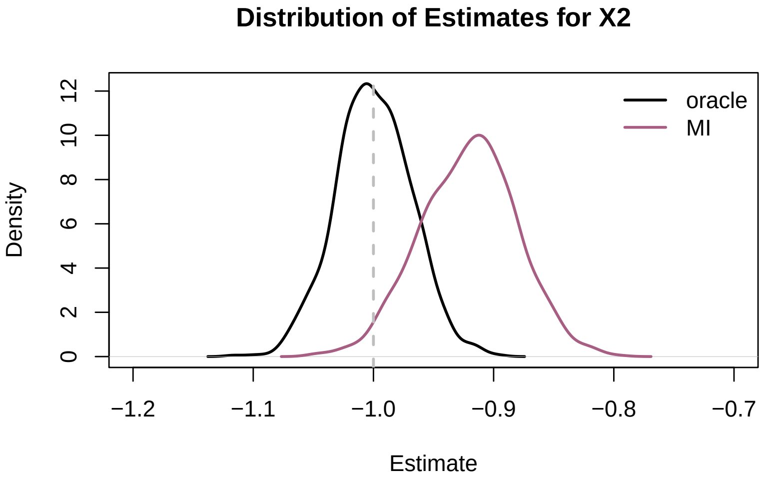
Results: MI under covers!



Results: why?

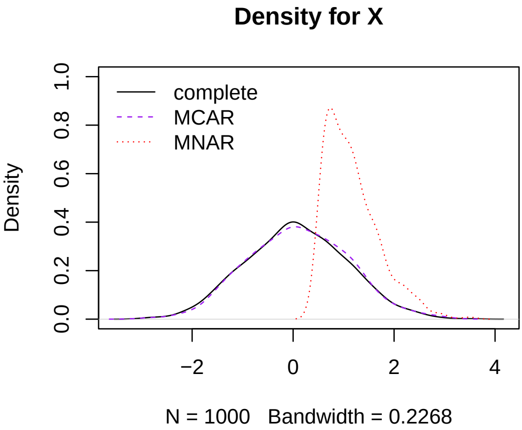

Distribution of Estimates for X1
Density
Estimate
oracle
MI
0
1
2
3
4
5
6
1.0
1.5
2.0
2.5

Results: why?


Distribution of Estimates for X2
oracle
MI
Density
0
2
4
6
8
10
12
−1.2
−1.1
−1.0
−0.9
−0.8
−0.7
Estimate

How do we think of missing data?

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



Missing NOT at Random

- Systematic nonrandomness:

$$X_{\text{mis}} \not\perp\!\!\!\perp R \mid X_{\text{obs}}$$

- Observed values are not enough to learn the imputation model.
- 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.

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.
- 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