Missing Data Imputation

5 maj

Outline

1. Instrumental Variables
2. Missing data
3. Single imputation methods
4. Multiple imputation

Missing data

- What is it?

Missing data

- What is it?
- Why are data missing?

Missing data

- What is it?
- Why are data missing?
- How often do we encounter missing data?

An Exercise

- Find your 3rd assignment for this course
- What was the sample size for your model?

Wooldridge

- Wooldridge only mentions missing data once:

> If the data are missing at random, then the size of the random sample available from the population is simply reduced. Although this makes the estimators less precise, it does not introduce any bias [...] There are ways to use the information on observations where only some variables are missing, but this is not often done in practice. The improvement in the estimators is usually slight, while the methods are somewhat complicated. In most cases, we just ignore the observations that have missing information.

Wooldridge p.314

Mising Data In Practice

- Ideally we don't have any missing data
- Since this is never the case we have to respond to it

Effects of Missing Data

1. Statistical efficiency
2. Comparability of analyses
3. Representativeness
4. Scale construction
5. Causal inference

Scaling

- We haven't talked a lot about scaling in this class
- How do we build scales with missing values?

An Example

- We often use simple additive scales

An Example

- We often use simple additive scales

Case	Item 1	Item 2	Item 3	Sum
A	1	2	1	?
B	1	.	3	?
C	.	1	1	?
D	2	1	2	?
E	1	.	.	?
F	.	.	.	?

Two Default Approaches

1. Complete Case Analysis
    - i.e., listwise deletion
2. Available Case Analysis
    - This is our default

Available Case Analysis

- Software default
- Remove any observations with missing values for a given analysis
- Benefits
    - Easy
    - Uses as much data as possible for any given analysis
- Consequences
    - Loss of information
    - Analyses are not comparable to one another
    - Loss of sample representativeness?

Complete Case Analysis

- Listwase or case deletion
- Remove any observations with missing values before all analyses
- Benefits
  - Easy
  - All analyses are comparable to one another
- Consequences
  - Loss of information (more than available case)
  - Loss of sample representativeness?

What can we do about missing data?

- Use default strategies
- Impute missing values once
- Run analyses multiple times with different imputations

Missing Data Assumptions

1. Missing Completely At Random (MCAR)
2. Missing At Random (MAR; Ignorable)
3. Nonignorable (NI)

Missing Data Assumptions

1. Missing Completely At Random (MCAR)
2. Missing At Random (MAR; Ignorable)
3. Nonignorable (NI)
4. Censoring

Examples?

1. Missing Completely At Random (MCAR)

Examples?

1. Missing Completely At Random (MCAR)
2. Missing At Random (MAR; Ignorable)

Examples?

1. Missing Completely At Random (MCAR)
2. Missing At Random (MAR; Ignorable)
3. Nonignorable (NI)

MCAR

1. Statistical efficiency
2. Comparability of analyses
3. Representativeness
4. Scale construction
5. Causal inference

MAR/Ignorable

1. Statistical efficiency
2. Comparability of analyses
3. Representativeness
4. Scale construction
5. Causal inference

NI

1. Statistical efficiency
2. Comparability of analyses
3. Representativeness
4. Scale construction
5. Causal inference

Questions?

Outline

1. Instrumental Variables
2. Missing data
3. Single imputation methods
4. Multiple imputation

Single imputation

- Impute missing values once
- Benefits
  - Increases statistical efficiency
  - Comparable analyses
  - Easy (depends on technique)
  - Preserve representativeness?
- Consequences
  - Bias (depends on technique)
  - Analyses do not reflect uncertainty due to missingness

Single Imputation Methods

- Single value (e.g. zero, mean)
- Random value
- Inferred value
- Hot deck
- Regression

Regression Imputation

- Fill in missing values using fitted values from a regression model
- We can use any variable in this model
    - There is no need to respect causal ordering
- Should we impute missing outcome values?

Questions?

Outline

1. Instrumental Variables
2. Missing data
3. Single imputation methods
4. Multiple imputation

Multiple Imputation (MI)

1. Engage a single imputation method
2. Repeat this method multiple times to create multiple imputed datasets
3. Run analysis on each imputed dataset
4. Combine estimates and calculate combined variances

Multiple Imputation (MI)

- Benefits
    - Increases statistical efficiency
    - Account for uncertainty due to missingness
    - Comparable analyses
    - Better than single imputation if MAR
- Consequences
    - Somewhat computationally expensive
    - Challenging with large datasets
    - Software may not, by default, handle all statistical tests in an MI framework

Multiple Imputation (MI)

- Easy if missingness only on one variable
- More complicated if missingness on multiple variables
- Still not very commonly used in political science
- Doesn't solve NI missingness

Aggregating MI Results

1. Run analysis on each imputed dataset
2. Each analysis produces a test statistic $\hat{\beta}_m$
3. Overall test statistic is $\Sigma_1^M \hat{\beta}_m$
4. Calculating variance is a sum of two things:
    - within-imputation variance
    - between-imputation variance

Variance of MI Estimates

- Within-imputation variance:

  $Within = \frac{1}{m} \Sigma_1^M \hat{V}_m$

- Between-imputation variance:

  $Between = \frac{1}{m-1} \Sigma_1^M (\hat{\beta}_m - \hat{\beta})^2$

- Total variance is:

  $V_\beta = Within + (1 + \frac{1}{m}) Between$

An Example

- What is the effect of university education on an individuals' political tolerance?
  - University education is a binary indicator
  - Tolerance is on an 11-point scale
- Missingness in various covariates
- Multiply impute missing values
- On each imputed dataset, we regress Tolerance on Education + Controls
- Our test statistic is $\beta_{E}duc$

An Example

- Here are some MI results:

Dataset	β_{Educ}	SE_{β}
1	4.32	0.95
2	4.15	1.16
3	4.86	0.83
4	3.98	1.04
5	4.50	0.91

- What is the overall β_{Educ} ?
- What is the overall SE_{β} ?

Questions?

Summary

- Missing data is a common problem
- Multiple strategies for coping with it
- Some better than others
- For exam and real life, be prepared to justify your strategy for addressing missing data