Introduction to missing covariate data

Sandrah P. Eckel
eckel@usc.edu

Background slides inspired by Tom Louis’ talk
“Missing Data/Measurement Error”
Motivation

• Missing data are inevitable
  ▪ Especially in large cohort studies with many predictors
• Care needed to make valid inferences for original target population
• Goal: discuss approaches to quantify associations of interest when missing
  ▪ predictor of interest
  ▪ confounder
  ▪ effect modifier
(cross-sectional)
Missing data mechanisms

1. Missing Completely at Random (MCAR)
   \( \text{Pr}(\text{missing}) \) unrelated to process under study

2. Missing at Random (MAR)
   \( \text{Pr}(\text{missing}) \) depends only on observed data

3. Not Missing at Random (NMAR)
   \( \text{Pr}(\text{missing}) \) depends on both observed and unobserved data

Validity of an analysis depends on the missing data mechanism

Little R, Rubin D. Statistical analysis with missing data. 2002.
Missing data mechanisms: examples

MCAR: unrelated to process studied

MAR: depends on observed data

NMAR: depends on unobserved
Common methods for missing data

- Complete case
- Missing indicator
- Single imputation
- Multiple imputation
- … and many others
Complete case analysis

• Common default method

Implementation
• Usually automatic
• Good practice to formally subset analysis dataset to complete cases

• \texttt{R} > \texttt{ccdat <- subset(dat, complete.cases(dat))}

Properties
• MCAR data: inferences valid, but inefficient
• Not MCAR data: inferences may be biased
  ▪ Study population = those with complete data
Missing Indicator Method

Implementation

• Categorical $X$
  ▪ Create an additional “missing” category and model as usual

• Continuous $X$
  ▪ Create new variable $X_{\text{fill}}$: missing values of $X$ set to a constant (0)
  ▪ Create an indicator variable for $X$ missing
  ▪ Include $X_{\text{fill}}$ and indicator in regression model

Properties

• Maintains sample size, treats missing data differently

• Controversy: some recommend strongly against it
  ▪ “…can exhibit severe bias even when the data are missing completely at random” – Greenland and Finkle, 1995
  ▪ Do we encounter this problem in practice?
Single Imputation

Implementation
1. Fill in missing values with imputed data
   - Mean, predicted values from regression, etc..
2. Apply standard analysis to complete data

Properties
• Underestimates uncertainty
  - Assumes imputed values are “true”
• Usually recommended to avoid this method
Multiple Imputation

Implementation
1. Fill in missing values with imputed data
   - Predicted value from regression + error, etc..
2. Repeat m times to create m complete data sets
   - m usually small (5-10)
3. Apply standard analysis to each data set
4. Combine estimates and inferences

Properties
• Acknowledges uncertainty in imputed data

Rubin D. Multiple Imputation for Nonresponse in Surveys. 1987
Multiple Imputation: in diagram

Standard complete data analysis

Original data: Some missing

Imputed data 1

Imputed data 2

Imputed data 3

Imputed data 4

Imputed data 5

Results 1

Results 2

Results 3

Results 4

Results 5

Pooled Results

Standard complete data analysis
Multiple Imputation: formulas and R

Pooled estimate and inference

\[ \bar{\beta} = \frac{1}{m} \sum_{k=1}^{m} \hat{\beta}^{(k)} \]

\[ \text{Var}(\bar{\beta}) = \frac{1}{m} \sum_{k=1}^{m} \text{Var}(\hat{\beta}^{(k)}) + \left(1 + \frac{1}{m}\right) \frac{1}{m-1} \sum_{k=1}^{m} \left(\hat{\beta}^{(k)} - \bar{\beta}\right)^2 \]

Within-imputation variance

Between-imputation variance

Sample R code (one of many MI implementations)

```r
library(mice)

datIncomplete <- data.frame(Y,X1,X2)
imp <- mice(datIncomplete, print=FALSE, m=5)
fit <- with(imp, lm(Y~X1*X2))
poolfit <- pool(fit)
```
Aside: missing Y

- For Y missing at random (MAR)
  - When $X$ complete:
    Incomplete cases provide no additional info to regression of Y on $X$
  - When some $X$ missing:
    Cases missing Y can provide additional info about missing $X$’s for cases with Y present
    - “Ignorable” with appropriate likelihood/Bayesian model
- For Y not MAR, analysis is more challenging
  - No simple solution - “Not ignorable”
  - One possible approach: sensitivity analysis
    - Dan Scharfstein’s work

Little, 1992. Regression with missing X’s: a review
Subset of Children’s Health Study Data

- N = 1541 (Cohort E, year 6)
  - Subset with complete eNO data (first eNO test)
- Outcome: Exhaled Nitric Oxide
  - log(eno) since right skewed
- Predictors:
  - Age: 9.9-13.1 years, no missing
  - Asthma: 13% positive, 8% missing
  - Allergy: 30% current, no missing
  - Second hand smoke:
    - 3% positive, 6% missing
- Following examples use a random subset of 1000 children out of the 1392 with complete data
Impose known missing data mechanisms on subset of complete CHS data

Missing

- Predictor of interest
- Confounder
- Effect modifier

Missing data mechanism is known, but “true” model for data is unknown
Single predictor of interest: age

\[ E(Y) = \beta_0 + \beta_1 \text{age} \]

|            | Estimate | SE  | t-value | Pr(>|t|) |
|------------|----------|-----|---------|----------|
| Intercept  | 1.69     | 0.40| 4.19    | 3.1e-05  *** |
| age        | 0.08     | 0.04| 2.14    | 0.032    * |
**Age missing randomly (MCAR)**

**Complete case analysis**

\[ E(Y) = \beta_0 + \beta_1 \text{age} \]
Age missing randomly (MCAR)
Missing Indicator for age

$$E(Y) = \beta_0 + \beta_1 age^* + \beta_2 I(age = \text{missing})$$

age*: set to 0 when age missing
Age missing randomly (MCAR)

Single Imputation: mean (\(X\))

\[ E(Y) = \beta_0 + \beta_1 \text{age} \]
Age missing randomly (MCAR)

Multiple Imputation (m=10): using Y and other X’s

\[ E(Y) = \beta_0 + \beta_1 \text{age} \]
Age missing randomly (MCAR)
1000 simulations

Bias

Proportion missing Age

SE

Proportion missing Age
Age missing dependent on asthma (MAR)
Age missing dependent on asthma (MAR)
1000 simulations

Bias

SE

Complete Case
Missing Indicator
Single Imputation
Multiple Imputation
Age missing dependent on age (NMAR)
Age missing dependent on age (NMAR)

1000 simulations

Bias

Proportion missing Age

SE

Proportion missing Age

Complete Case
Missing Indicator
Single Imputation
Multiple Imputation
Potential confounding of asthma-eNO association by age

Complete data:
No evidence that $\beta_1$ is confounded by age

\[ E(Y) = \beta_0 + \beta_1 \text{asthma} \]

| Estimate  | Std. Error | t value     | Pr(>|t|)   |
|-----------|------------|-------------|------------|
| (Intercept) | 2.46179    | 0.02367     | 103.987 < 2e-16 *** |
| asthma    | 0.49936    | 0.06217     | 8.032 2.69e-15 *** |

\[ E(Y) = \beta_0 + \beta_1 \text{asthma} + \beta_2 \text{age} \]

| Estimate  | Std. Error | t value     | Pr(>|t|)   |
|-----------|------------|-------------|------------|
| (Intercept) | 1.77887    | 0.40143     | 4.431 1.04e-05 *** |
| asthma    | 0.49965    | 0.06211     | 8.044 2.45e-15 *** |
| age       | 0.06040    | 0.03544     | 1.704 0.0887 . |

Incomplete data:
Estimate $\beta_1$ when some age values missing
Potential confounder missing

\[ E(Y) = \beta_0 + \beta_1 \text{asthma} + \beta_2 \text{age} \]
Potential modification of asthma-eNO association by current allergy

Complete data:
Asthma coefficient is larger in those with current allergy \((\beta_1 + \beta_3)\) than for those without \((\beta_1)\)

\[
E(Y) = \beta_0 + \beta_1 \text{asthma} + \beta_2 \text{allergy} + \beta_3 \text{asthma} \times \text{allergy}
\]

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 2.38191 | 0.02706 | 88.039 | < 2e-16 *** |
| asthma | 0.27454 | 0.09677 | 2.837 | 0.00464 ** |
| allergyCur | 0.29696 | 0.05216 | 5.693 | 1.64e-08 *** |
| asthma:allergyCur | 0.18327 | 0.12777 | 1.434 | 0.15178 |

Incomplete data:
What happens to estimate of \(\beta_1\) (asthma coefficient for those without current allergy) when some current allergy values are missing?
Potential effect modifier missing

\[ E(Y) = \beta_0 + \beta_1 \text{asthma} + \beta_2 \text{allergy} + \beta_3 \text{asthma} \times \text{allergy} \]
Real missing data mechanism

\[ E(Y) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{asthma} + \beta_3 \text{allergy} + \beta_4 \text{SHS} \]

Estimates and inference similar for all 4 analysis techniques
Summary: missing data analysis methods

Good practice to

• Report and summarize missing data
• Report analysis method used to handle missing data
• Assess sensitivity of key findings to analysis method choice

General comments

• Differences between analysis methods may not be that large in some practical situations?
• Still need to fit “appropriate” models, regardless of analysis method

Discussion

• What is your usual approach?
• Should there be a ‘recommended default’ approach?
Missing data treatment in practice
- Burton and Altman, 2004

100 articles reviewed

No missing covariate data in 15 articles

13 had availability of data as an inclusion criteria

Two were randomised trials

Missing covariate data in 81 articles

Unknown whether any missing covariate data in four articles

“The methods used to handle incomplete covariates were obtainable in 32 of the 81 articles with known missing data and the most commonly reported approaches were complete case and available case analysis.”

Proposed guidelines - Burton and Altman, 2004
(1 of the 100 papers satisfied these guidelines)

Quantification of completeness of covariate data
- If availability of data is an inclusion criterion, specify the number of cases excluded for this reason
- Provide the total number of eligible cases and the number with complete data
- Report the frequency of missing data for every variable considered. If there is only a small amount
  of overall missingness (e.g. >90% of cases with complete data), then the number of incomplete
  variables and the maximum amount of missingness in any variable are sufficient

Approaches for handling missing covariate data
- Provide sufficient details of the methods adopted to handle missing covariate data for all
  incomplete covariates
- Give appropriate references for any imputation method used
- For each analysis, specify the number of cases included and the associated number of events

Exploration of the missing data
- Discuss any known reasons for missing covariate data
- Present the results of any comparisons of characteristics between the cases with without
  missing data

Age missing dependent on Y (MAR)
Age missing dependent on Y (MAR)
1000 simulations

Bias

SE

Proportion missing Age

Proportion missing Age