Statistical Methods for Analysis with Missing Data

Lecture 3: naïve methods: complete-case analysis and imputation

Mauricio Sadinle

Department of Biostatistics

UNIVERSITY of WASHINGTON

Previous Lecture

Universe of missing-data mechanisms:

- MCAR: $p(R = r \mid z) = p(R = r)$
  - Unreasonable in most cases
- MAR: $p(R = r \mid z) = p(R = r \mid z_{(r)})$
  - Hard to digest, in general
  - $R \perp\!\!\!\perp Z_1 \mid Z_2,$ if Z_2 fully observed
- MNAR: $p(R = r \mid z) \neq p(R = r \mid z_{(r)})$
  - Most realistic, but hard to handle

Today's Lecture

Naïve or ad-hoc methods

- Complete-case / available-case analyses
- Different types of (single) imputation

Reading: Ch. 2, of Davidian and Tsiatis

Naïve or Ad-Hoc Methods

- Motivation: we know how to run analyses with complete (rectangular) datasets
- Idea: somehow "fix" the dataset so that the analysis for complete data can be run

Outline

Outline

Complete-Case Analysis

- Idea: ignore observations with missingness, run intended analysis with remaining data

Complete-Case Analysis

Gender	Age	Income	...
F	25	60,000	...
~~M~~	~~?~~	~~?~~	~~...~~
~~?~~	~~51~~	~~?~~	~~...~~
~~F~~	~~?~~	~~150,300~~	
...	...	...	...

Assumption for Complete-Case Analysis

Complete-case analysis implicitly assumes

$$p(z) = p(z \mid R = 1_K) \tag{1}$$

where 1_K represents a vector $(1, 1, \ldots, 1)$ of length K

- By Bayes' theorem

$$p(z \mid R = 1_K) = \frac{p(R = 1_K \mid z)p(z)}{p(R = 1_K)}$$

- Therefore, (1) is equivalent to

$$p(R = 1_K \mid z) = p(R = 1_K)$$

- This doesn't require any assumptions on $p(R = r \mid z)$ for $r \neq 1_K$
- MCAR ($Z \perp\!\!\!\perp R$) is a sufficient condition for (1)

Assumption for Complete-Case Analysis

Complete-case analysis implicitly assumes

$$p(z) = p(z \mid R = 1_K) \tag{1}$$

where 1_K represents a vector $(1, 1, \ldots, 1)$ of length K

- By Bayes' theorem

$$p(z \mid R = 1_K) = \frac{p(R = 1_K \mid z)p(z)}{p(R = 1_K)}$$

- Therefore, (1) is equivalent to

$$p(R = 1_K \mid z) = p(R = 1_K)$$

- This doesn't require any assumptions on $p(R = r \mid z)$ for $r \neq 1_K$
- MCAR ($Z \perp\!\!\!\perp R$) is a sufficient condition for (1)

Assumption for Complete-Case Analysis

Complete-case analysis implicitly assumes

$$p(z) = p(z \mid R = 1_K) \tag{1}$$

where 1_K represents a vector $(1, 1, \ldots, 1)$ of length K

- By Bayes' theorem

$$p(z \mid R = 1_K) = \frac{p(R = 1_K \mid z)p(z)}{p(R = 1_K)}$$

- Therefore, (1) is equivalent to

$$p(R = 1_K \mid z) = p(R = 1_K)$$

- This doesn't require any assumptions on $p(R = r \mid z)$ for $r \neq 1_K$
- MCAR ($Z \perp\!\!\!\perp R$) is a sufficient condition for (1)

Assumption for Complete-Case Analysis

Complete-case analysis implicitly assumes

$$p(z) = p(z \mid R = 1_K) \tag{1}$$

where 1_K represents a vector $(1, 1, \ldots, 1)$ of length K

- By Bayes' theorem

$$p(z \mid R = 1_K) = \frac{p(R = 1_K \mid z)p(z)}{p(R = 1_K)}$$

- Therefore, (1) is equivalent to

$$p(R = 1_K \mid z) = p(R = 1_K)$$

- This doesn't require any assumptions on $p(R = r \mid z)$ for $r \neq 1_K$

- MCAR ($Z \perp\!\!\!\perp R$) is a sufficient condition for (1)

Assumption for Complete-Case Analysis

Complete-case analysis implicitly assumes

$$p(z) = p(z \mid R = 1_K) \tag{1}$$

where 1_K represents a vector $(1, 1, \ldots, 1)$ of length K

- By Bayes' theorem

$$p(z \mid R = 1_K) = \frac{p(R = 1_K \mid z)p(z)}{p(R = 1_K)}$$

- Therefore, (1) is equivalent to

$$p(R = 1_K \mid z) = p(R = 1_K)$$

- This doesn't require any assumptions on $p(R = r \mid z)$ for $r \neq 1_K$
- MCAR ($Z \perp\!\!\!\perp R$) is a sufficient condition for (1)

Complete-Case Analysis is Wasteful/Inefficient

Clearly, there can be a huge waste of information

- Observed data with response patterns $r \neq 1_K$ should be informative about the distribution of $Z_{(r)}$, which is informative about the distribution of Z

$$p(z_{(r)}) = \int p(z)\ dz_{(\bar{r})}, \quad r \in \{0,1\}^K$$

- We might end up with very little data
  - Say the $R_1, \dots, R_K \overset{i.i.d.}{\sim} \text{Bernoulli}(\pi)$
    - $p(R = 1_K) = \pi^K \quad \overset{K \to \infty}{\longrightarrow} \quad 0$

Complete-Case Analysis is Wasteful/Inefficient

Clearly, there can be a huge waste of information

- Observed data with response patterns $r \neq 1_K$ should be informative about the distribution of $Z_{(r)}$, which is informative about the distribution of Z

$$p(z_{(r)}) = \int p(z)\ dz_{(\bar{r})}, \quad r \in \{0,1\}^K$$

- We might end up with very little data
  - Say the $R_1, \ldots, R_K \overset{i.i.d.}{\sim}$ Bernoulli(π)
    - $p(R = 1_K) = \pi^K \quad \xrightarrow{K \to \infty} \quad 0$

Example: Estimating a Mean

We'll see an alternative presentation of Example 1 in Section 1.4 of Davidian and Tsiatis

- $\{(Y_i, R_i)\}_{i=1}^n \overset{i.i.d.}{\sim} F$
- Y_i: numeric variable for individual i
- R_i: indicator of Y_i being observed
- If Y_i was always observed, we could estimate the mean of Y, $\mu = E(Y)$, as

$$\hat{\mu}^{full} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

Example: Estimating a Mean

With missing data, we could use the complete cases

$$\hat{\mu}^{cc} = \frac{\sum_{i=1}^{n} Y_i R_i}{\sum_{i=1}^{n} R_i}$$

Is this any good?

HW1: show that the following holds

$$E(\hat{\mu}^{cc}) = E(Y \mid R = 1)$$

for all sample sizes, provided that at least one Y_i is observed.

Hint: write $E(\hat{\mu}^{cc}) = E\left[E\left(\frac{\sum_{i=1}^{n} Y_i R_i}{\sum_{i=1}^{n} R_i} \mid R_1, \ldots, R_n\right)\right]$

Example: Estimating a Mean

With missing data, we could use the complete cases

$$\hat{\mu}^{cc} = \frac{\sum_{i=1}^{n} Y_i R_i}{\sum_{i=1}^{n} R_i}$$

Is this any good?

HW1: show that the following holds

$$E(\hat{\mu}^{cc}) = E(Y \mid R = 1)$$

for all sample sizes, provided that at least one Y_i is observed.

Hint: write $E(\hat{\mu}^{cc}) = E\left[E\left(\frac{\sum_{i=1}^{n} Y_i R_i}{\sum_{i=1}^{n} R_i} \mid R_1, \ldots, R_n\right)\right]$

Example: Estimating a Mean

With missing data, we could use the complete cases

$$\hat{\mu}^{cc} = \frac{\sum_{i=1}^{n} Y_i R_i}{\sum_{i=1}^{n} R_i}$$

Is this any good?

HW1: show that the following holds

$$E(\hat{\mu}^{cc}) = E(Y \mid R = 1)$$

for all sample sizes, provided that at least one Y_i is observed.

Hint: write $E(\hat{\mu}^{cc}) = E\left[E\left(\frac{\sum_{i=1}^{n} Y_i R_i}{\sum_{i=1}^{n} R_i} \mid R_1, \ldots, R_n\right)\right]$

Example: Estimating a Mean

$$E(\hat{\mu}^{cc}) = E(Y \mid R = 1)$$

Therefore

- Complete-case estimator of the mean requires assuming

$$E(Y) = E(Y \mid R = 1)$$

- In particular, valid under MCAR
- Otherwise, $\hat{\mu}^{cc}$ is not valid for μ, as it estimates the wrong quantity
- HW1: if $p(R = 1 \mid y)$ is an increasing function of y, show that

$$E(Y \mid R = 1) > E(Y)$$

Example: Estimating a Mean

$$E(\hat{\mu}^{cc}) = E(Y \mid R = 1)$$

Therefore

- Complete-case estimator of the mean requires assuming

$$E(Y) = E(Y \mid R = 1)$$

- In particular, valid under MCAR
- Otherwise, $\hat{\mu}^{cc}$ is not valid for μ, as it estimates the wrong quantity
- HW1: if $p(R = 1 \mid y)$ is an increasing function of y, show that

$$E(Y \mid R = 1) > E(Y)$$

Outline

Available-Case Analysis

Sometimes what we need to estimate doesn't really require a "rectangular" dataset

- If you can, just use whatever data are available for computing what you need
- Davidian and Tsiatis talk about generalized estimating equations (GEEs) and their Example 3 in Section 1.4 (we'll cover this when we get to Chapter 5)
- K normal random variables: under some missing-data assumption, it seems we could still obtain a good estimate of the distribution as it only depends on univariate and bivariate quantities (means, variances, covariances)

Available-Case Analysis

Sometimes what we need to estimate doesn't really require a "rectangular" dataset

- If you can, just use whatever data are available for computing what you need
- Davidian and Tsiatis talk about generalized estimating equations (GEEs) and their Example 3 in Section 1.4 (we'll cover this when we get to Chapter 5)
- K normal random variables: under some missing-data assumption, it seems we could still obtain a good estimate of the distribution as it only depends on univariate and bivariate quantities (means, variances, covariances)

Available-Case Analysis

Sometimes what we need to estimate doesn't really require a "rectangular" dataset

- If you can, just use whatever data are available for computing what you need
- Davidian and Tsiatis talk about generalized estimating equations (GEEs) and their Example 3 in Section 1.4 (we'll cover this when we get to Chapter 5)
- K normal random variables: under some missing-data assumption, it seems we could still obtain a good estimate of the distribution as it only depends on univariate and bivariate quantities (means, variances, covariances)

Example of Available-Case Analysis

- Say the data are
  - $Z_i = (Y_{i1}, \ldots, Y_{iK})$
  - $R_i = (R_{i1}, \ldots, R_{iK})$
- Available-case estimators:

$$\hat{\mu}_j^{ac} = \frac{\sum_{i=1}^n Y_{ij} R_{ij}}{\sum_{i=1}^n R_{ij}}, \quad j = 1, \ldots, K$$

$$\hat{\sigma}_{jk}^{ac} = \frac{\sum_{i=1}^n (Y_{ij} - \hat{\mu}_j^{ac})(Y_{ik} - \hat{\mu}_k^{ac}) R_{ij} R_{ik}}{\sum_{i=1}^n R_{ij} R_{ik} - 1}; \quad j, k = 1, \ldots, K$$

- Better than complete-case analysis
- Valid under MCAR, but what are the minimal assumptions on the missing-data mechanism for this to be valid?

Example of Available-Case Analysis

- Say the data are
  - $Z_i = (Y_{i1}, \ldots, Y_{iK})$
  - $R_i = (R_{i1}, \ldots, R_{iK})$
- Available-case estimators:

$$\hat{\mu}_j^{ac} = \frac{\sum_{i=1}^n Y_{ij} R_{ij}}{\sum_{i=1}^n R_{ij}}, \quad j = 1, \ldots, K$$

$$\hat{\sigma}_{jk}^{ac} = \frac{\sum_{i=1}^n (Y_{ij} - \hat{\mu}_j^{ac})(Y_{ik} - \hat{\mu}_k^{ac}) R_{ij} R_{ik}}{\sum_{i=1}^n R_{ij} R_{ik} - 1}; \quad j, k = 1, \ldots, K$$

- Better than complete-case analysis
- Valid under MCAR, but what are the minimal assumptions on the missing-data mechanism for this to be valid?

Example of Available-Case Analysis

- Say the data are
  - $Z_i = (Y_{i1}, \ldots, Y_{iK})$
  - $R_i = (R_{i1}, \ldots, R_{iK})$
- Available-case estimators:

$$\hat{\mu}_j^{ac} = \frac{\sum_{i=1}^n Y_{ij} R_{ij}}{\sum_{i=1}^n R_{ij}}, \quad j = 1, \ldots, K$$

$$\hat{\sigma}_{jk}^{ac} = \frac{\sum_{i=1}^n (Y_{ij} - \hat{\mu}_j^{ac})(Y_{ik} - \hat{\mu}_k^{ac}) R_{ij} R_{ik}}{\sum_{i=1}^n R_{ij} R_{ik} - 1}; \quad j, k = 1, \ldots, K$$

- Better than complete-case analysis
- Valid under MCAR, but what are the minimal assumptions on the missing-data mechanism for this to be valid?

Example of Available-Case Analysis

- Say the data are
  - $Z_i = (Y_{i1}, \dots, Y_{iK})$
  - $R_i = (R_{i1}, \dots, R_{iK})$
- Available-case estimators:

$$\hat{\mu}_j^{ac} = \frac{\sum_{i=1}^n Y_{ij} R_{ij}}{\sum_{i=1}^n R_{ij}}, \quad j = 1, \dots, K$$

$$\hat{\sigma}_{jk}^{ac} = \frac{\sum_{i=1}^n (Y_{ij} - \hat{\mu}_j^{ac})(Y_{ik} - \hat{\mu}_k^{ac}) R_{ij} R_{ik}}{\sum_{i=1}^n R_{ij} R_{ik} - 1}; \quad j, k = 1, \dots, K$$

- Better than complete-case analysis
- Valid under MCAR, but what are the minimal assumptions on the missing-data mechanism for this to be valid?

Complete-Case and Available-Case Analysis

The moral:

- Complete-case analysis is wasteful and, most likely, invalid
- Available-case analysis is better, but still requires MCAR or possibly a weaker assumption depending on what we need to compute

Outline

Imputation

- Idea: plug something "reasonable" into the holes of the dataset, then run intended analysis with completed data

Imputation

Gender	Age	Income	...
F	25	60,000	...
M	? 20	? 30,000	...
? M	51	? 70,000	...
F	? 60	150,300	
...	...	...	...

Outline

Mean Imputation

- Numeric variables
  - Impute mean of observed values
  - Corresponds to imputing an estimate of $E(Y_j \mid R_j = 1)$, $j = 1, \ldots, K$
  - Leads to valid point estimates of means under MCAR
  - Underestimates true variance of estimators

Mean Imputation

Say the data are

- $\{(Z_i, R_i)\}_{i=1}^n \overset{i.i.d.}{\sim} F$
- $Z_i = (Y_{i1}, \ldots, Y_{iK})$
- $R_i = (R_{i1}, \ldots, R_{iK})$

Mean imputation:

- Compute

$$\hat{\mu}_j^1 = \frac{\sum_{i=1}^n Y_{ij} R_{ij}}{\sum_{i=1}^n R_{ij}}, \quad j = 1, \ldots, K$$

- Impute Y_{ij} with $\hat{\mu}_j^1$ whenever $R_{ij} = 0$
- Run your analysis as if your data were fully observed

Mean Imputation

Say the data are

- $\{(Z_i, R_i)\}_{i=1}^{n} \overset{i.i.d.}{\sim} F$
- $Z_i = (Y_{i1}, \ldots, Y_{iK})$
- $R_i = (R_{i1}, \ldots, R_{iK})$

Mean imputation:

- Compute

$$\hat{\mu}_j^1 = \frac{\sum_{i=1}^{n} Y_{ij} R_{ij}}{\sum_{i=1}^{n} R_{ij}}, \quad j = 1, \ldots, K$$

- Impute Y_{ij} with $\hat{\mu}_j^1$ whenever $R_{ij} = 0$
- Run your analysis as if your data were fully observed

Mean Imputation

Say the data are

- $\{(Z_i, R_i)\}_{i=1}^{n} \overset{i.i.d.}{\sim} F$
- $Z_i = (Y_{i1}, \ldots, Y_{iK})$
- $R_i = (R_{i1}, \ldots, R_{iK})$

Mean imputation:

- Compute
$$\hat{\mu}_j^1 = \frac{\sum_{i=1}^{n} Y_{ij} R_{ij}}{\sum_{i=1}^{n} R_{ij}}, \quad j = 1, \ldots, K$$
- Impute Y_{ij} with $\hat{\mu}_j^1$ whenever $R_{ij} = 0$
- Run your analysis as if your data were fully observed

Mean Imputation

Say the data are

- $\{(Z_i, R_i)\}_{i=1}^{n} \overset{i.i.d.}{\sim} F$
- $Z_i = (Y_{i1}, \dots, Y_{iK})$
- $R_i = (R_{i1}, \dots, R_{iK})$

Mean imputation:

- Compute

$$\hat{\mu}_j^1 = \frac{\sum_{i=1}^{n} Y_{ij} R_{ij}}{\sum_{i=1}^{n} R_{ij}}, \quad j = 1, \dots, K$$

- Impute Y_{ij} with $\hat{\mu}_j^1$ whenever $R_{ij} = 0$
- Run your analysis as if your data were fully observed

Mean Imputation

Age	*Income*
25	60,000
?	?
51	?
?	150,300
⋮	⋮

$\Longrightarrow$

Age	*Income*
25	60,000
$\hat{\mu}^1_{Age}$	$\hat{\mu}^1_{Income}$
51	$\hat{\mu}^1_{Income}$
$\hat{\mu}^1_{Age}$	150,300
⋮	⋮

Example: Estimating a Mean

- Estimating a mean after mean imputation corresponds to using the estimator
$$\hat{\mu}_j^{mimp} = \frac{1}{n}\sum_{i=1}^{n}[Y_{ij}R_{ij} + \hat{\mu}_j^1(1 - R_{ij})]$$
- $\hat{\mu}_j^{mimp}$ is the mean of the imputed data, so its naïvely estimated variance is
$$\hat{V}_{\text{naïve}}(\hat{\mu}_j^{mimp}) = \hat{V}_{\text{naïve}}(Y_j)/n$$
where
$$\hat{V}_{\text{naïve}}(Y_j) = \frac{1}{n-1}\sum_{i=1}^{n}[R_{ij}(Y_{ij} - \hat{\mu}_j^{mimp})^2 + (1 - R_{ij})(\hat{\mu}_j^1 - \hat{\mu}_j^{mimp})^2]$$
- HW1: show that $\hat{\mu}_j^{mimp} = \hat{\mu}_j^1$

Example: Estimating a Mean

- Estimating a mean after mean imputation corresponds to using the estimator

$$\hat{\mu}_j^{mimp} = \frac{1}{n}\sum_{i=1}^{n}[Y_{ij}R_{ij} + \hat{\mu}_j^1(1 - R_{ij})]$$

- $\hat{\mu}_j^{mimp}$ is the mean of the imputed data, so its naïvely estimated variance is

$$\hat{V}_{\text{naïve}}(\hat{\mu}_j^{mimp}) = \hat{V}_{\text{naïve}}(Y_j)/n$$

where

$$\hat{V}_{\text{naïve}}(Y_j) = \frac{1}{n-1}\sum_{i=1}^{n}[R_{ij}(Y_{ij} - \hat{\mu}_j^{mimp})^2 + (1 - R_{ij})(\hat{\mu}_j^1 - \hat{\mu}_j^{mimp})^2]$$

- HW1: show that $\hat{\mu}_j^{mimp} = \hat{\mu}_j^1$

Example: Estimating a Mean

- Estimating a mean after mean imputation corresponds to using the estimator

$$\hat{\mu}_j^{mimp} = \frac{1}{n}\sum_{i=1}^{n}[Y_{ij}R_{ij} + \hat{\mu}_j^1(1 - R_{ij})]$$

- $\hat{\mu}_j^{mimp}$ is the mean of the imputed data, so its naïvely estimated variance is

$$\hat{V}_{\text{naïve}}(\hat{\mu}_j^{mimp}) = \hat{V}_{\text{naïve}}(Y_j)/n$$

where

$$\hat{V}_{\text{naïve}}(Y_j) = \frac{1}{n-1}\sum_{i=1}^{n}[R_{ij}(Y_{ij} - \hat{\mu}_j^{mimp})^2 + (1 - R_{ij})(\hat{\mu}_j^1 - \hat{\mu}_j^{mimp})^2]$$

- HW1: show that $\hat{\mu}_j^{mimp} = \hat{\mu}_j^1$

Example: Estimating a Mean

As a consequence, using the mean imputation method we:

- Underestimate the variance of each variable:

$$\hat{V}_{\text{naïve}}(Y_j) = \frac{1}{n-1} \sum_{i=1}^{n} R_{ij}(Y_{ij} - \hat{\mu}_j^1)^2$$

- Compare with an estimate based on the available cases:

$$\hat{V}^1(Y_j) = \frac{\sum_{i=1}^{n} R_{ij}(Y_{ij} - \hat{\mu}_j^1)^2}{\sum_{i=1}^{n} R_{ij} - 1}$$

- $\implies \hat{V}_{\text{naïve}}(Y_j) \leq \hat{V}^1(Y_j)$

Example: Estimating a Mean

As a consequence, using the mean imputation method we:

- Underestimate the variance of $\hat{\mu}_j^{mimp}$:

$$\hat{V}_{\text{naïve}}(\hat{\mu}_j^{mimp}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} R_{ij}(Y_{ij} - \hat{\mu}_j^1)^2$$

- Compare with an estimate based on the available cases:

$$\hat{V}^1(\hat{\mu}_j^{mimp}) = \frac{\sum_{i=1}^{n} R_{ij}(Y_{ij} - \hat{\mu}_j^1)^2}{(\sum_{i=1}^{n} R_{ij})(\sum_{i=1}^{n} R_{ij} - 1)}$$

- $\implies \hat{V}_{\text{naïve}}(\hat{\mu}_j^{mimp}) \leq \hat{V}^1(\hat{\mu}_j^{mimp})$
- HW1: comment on the implications of mean imputation for the construction of confidence intervals

Example: Estimating a Mean

As a consequence, using the mean imputation method we:

- Underestimate the variance of $\hat{\mu}_j^{mimp}$:

$$\hat{V}_{\text{naïve}}(\hat{\mu}_j^{mimp}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} R_{ij}(Y_{ij} - \hat{\mu}_j^1)^2$$

- Compare with an estimate based on the available cases:

$$\hat{V}^1(\hat{\mu}_j^{mimp}) = \frac{\sum_{i=1}^{n} R_{ij}(Y_{ij} - \hat{\mu}_j^1)^2}{(\sum_{i=1}^{n} R_{ij})(\sum_{i=1}^{n} R_{ij} - 1)}$$

- $\Longrightarrow \hat{V}_{\text{naïve}}(\hat{\mu}_j^{mimp}) \leq \hat{V}^1(\hat{\mu}_j^{mimp})$
- HW1: comment on the implications of mean imputation for the construction of confidence intervals

Outline

Mode Imputation

- Categorical variables
  - Impute mode of observed values
  - Artificially inflates frequency of mode
  - Leads to valid point estimates of marginal modes under MCAR
  - Underestimates true variance of estimators

Outline

Regression Imputation

- Regress one variable on others based on observed data, then impute predicted values from model
- Corresponds to imputing an estimate of $E(Y_j \mid y_{-j}, R = 1_K)$, where $y_{-j} = (y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_K)$
- Valid for means under MCAR
- Underestimates true variance of estimators
- Validity depends on model used for imputation

Regression Imputation

- Regress one variable on others based on observed data, then impute predicted values from model
- Corresponds to imputing an estimate of $E(Y_j \mid y_{-j}, R = 1_K)$, where $y_{-j} = (y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_K)$
- Valid for means under MCAR
- Underestimates true variance of estimators
- Validity depends on model used for imputation

Regression Imputation

- Regress one variable on others based on observed data, then impute predicted values from model
- Corresponds to imputing an estimate of $E(Y_j \mid y_{-j}, R = 1_K)$, where $y_{-j} = (y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_K)$
- Valid for means under MCAR
- Underestimates true variance of estimators
- Validity depends on model used for imputation

Regression Imputation

- Regress one variable on others based on observed data, then impute predicted values from model
- Corresponds to imputing an estimate of $E(Y_j \mid y_{-j}, R = 1_K)$, where $y_{-j} = (y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_K)$
- Valid for means under MCAR
- Underestimates true variance of estimators
- Validity depends on model used for imputation

Regression Imputation

- Regress one variable on others based on observed data, then impute predicted values from model
- Corresponds to imputing an estimate of $E(Y_j \mid y_{-j}, R = 1_K)$, where $y_{-j} = (y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_K)$
- Valid for means under MCAR
- Underestimates true variance of estimators
- Validity depends on model used for imputation

Example of Regression Imputation in Davidian and Tsiatis

- $Z = (Y_1, Y_2)$, baseline and follow-up, Y_1 always observed
- R indicator of response for Y_2
- Goal: to estimate $\mu_2 = E(Y_2)$
- Say we posit a linear model $E(Y_2 \mid y_1) = \beta_0 + \beta_1 y_1$
- Impute Y_{i2} with $\hat{Y}_{i2} = \hat{\beta}_0 + \hat{\beta}_1 Y_{i1}$ when $R_i = 0$, with $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained via least squares among complete cases
- The regression imputation estimator for μ_2 is

$$\hat{\mu}_2^{rimp} = \frac{1}{n} \sum_{i=1}^{n} [Y_{i2} R_i + \hat{Y}_{i2}(1 - R_i)]$$

- When is this valid? (when does $\hat{\mu}_2^{rimp} \overset{n \to \infty}{\longrightarrow} \mu_2$?)

Example of Regression Imputation in Davidian and Tsiatis

- $Z = (Y_1, Y_2)$, baseline and follow-up, Y_1 always observed
- R indicator of response for Y_2
- Goal: to estimate $\mu_2 = E(Y_2)$
- Say we posit a linear model $E(Y_2 \mid y_1) = \beta_0 + \beta_1 y_1$
- Impute Y_{i2} with $\hat{Y}_{i2} = \hat{\beta}_0 + \hat{\beta}_1 Y_{i1}$ when $R_i = 0$, with $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained via least squares among complete cases
- The regression imputation estimator for μ_2 is

$$\hat{\mu}_2^{rimp} = \frac{1}{n} \sum_{i=1}^{n} [Y_{i2} R_i + \hat{Y}_{i2}(1 - R_i)]$$

- When is this valid? (when does $\hat{\mu}_2^{rimp} \xrightarrow{n \to \infty} \mu_2$?)

Example of Regression Imputation in Davidian and Tsiatis

- $Z = (Y_1, Y_2)$, baseline and follow-up, Y_1 always observed
- R indicator of response for Y_2
- Goal: to estimate $\mu_2 = E(Y_2)$
- Say we posit a linear model $E(Y_2 \mid y_1) = \beta_0 + \beta_1 y_1$
- Impute Y_{i2} with $\hat{Y}_{i2} = \hat{\beta}_0 + \hat{\beta}_1 Y_{i1}$ when $R_i = 0$, with $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained via least squares among complete cases
- The regression imputation estimator for μ_2 is

$$\hat{\mu}_2^{rimp} = \frac{1}{n} \sum_{i=1}^{n} [Y_{i2} R_i + \hat{Y}_{i2}(1 - R_i)]$$

- When is this valid? (when does $\hat{\mu}_2^{rimp} \xrightarrow{n \to \infty} \mu_2$?)

Example of Regression Imputation in Davidian and Tsiatis

- $Z = (Y_1, Y_2)$, baseline and follow-up, Y_1 always observed
- R indicator of response for Y_2
- Goal: to estimate $\mu_2 = E(Y_2)$
- Say we posit a linear model $E(Y_2 \mid y_1) = \beta_0 + \beta_1 y_1$
- Impute Y_{i2} with $\hat{Y}_{i2} = \hat{\beta}_0 + \hat{\beta}_1 Y_{i1}$ when $R_i = 0$, with $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained via least squares among complete cases
- The regression imputation estimator for μ_2 is

$$\hat{\mu}_2^{rimp} = \frac{1}{n}\sum_{i=1}^{n}[Y_{i2}R_i + \hat{Y}_{i2}(1 - R_i)]$$

- When is this valid? (when does $\hat{\mu}_2^{rimp} \xrightarrow{n\to\infty} \mu_2$?)

Example of Regression Imputation in Davidian and Tsiatis

- $Z = (Y_1, Y_2)$, baseline and follow-up, Y_1 always observed
- R indicator of response for Y_2
- Goal: to estimate $\mu_2 = E(Y_2)$
- Say we posit a linear model $E(Y_2 \mid y_1) = \beta_0 + \beta_1 y_1$
- Impute Y_{i2} with $\hat{Y}_{i2} = \hat{\beta}_0 + \hat{\beta}_1 Y_{i1}$ when $R_i = 0$, with $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained via least squares among complete cases
- The regression imputation estimator for μ_2 is

$$\hat{\mu}_2^{rimp} = \frac{1}{n} \sum_{i=1}^{n} [Y_{i2} R_i + \hat{Y}_{i2}(1 - R_i)]$$

- When is this valid? (when does $\hat{\mu}_2^{rimp} \overset{n \to \infty}{\longrightarrow} \mu_2$?)

Example of Regression Imputation in Davidian and Tsiatis

Davidian and Tsiatis show that for $\hat{\mu}_2^{rimp} \xrightarrow{n\to\infty} \mu_2$ $(\hat{\mu}_2^{rimp} \xrightarrow{p} \mu_2)$ we need these two requirements to hold simultaneously:

- $E(Y_2 \mid y_1, R = 1) = E(Y_2 \mid y_1)$ (implied by MAR)
- $E(Y_2 \mid y_1)$ is correctly specified, i.e., there really exist β_0^* and β_1^* such that $E(Y_2 \mid y_1) = \beta_0^* + \beta_1^* y_1$

However, even if these two conditions hold, single imputation leads to underestimation of variances, as seen with mean imputation

Example of Regression Imputation in Davidian and Tsiatis

Davidian and Tsiatis show that for $\hat{\mu}_2^{rimp} \xrightarrow{n\to\infty} \mu_2$ $(\hat{\mu}_2^{rimp} \xrightarrow{p} \mu_2)$ we need these two requirements to hold simultaneously:

- $E(Y_2 \mid y_1, R = 1) = E(Y_2 \mid y_1)$ (implied by MAR)
- $E(Y_2 \mid y_1)$ is correctly specified, i.e., there really exist β_0^* and β_1^* such that $E(Y_2 \mid y_1) = \beta_0^* + \beta_1^* y_1$

However, even if these two conditions hold, single imputation leads to underestimation of variances, as seen with mean imputation

Example of Regression Imputation in Davidian and Tsiatis

Davidian and Tsiatis show that for $\hat{\mu}_2^{rimp} \overset{n\to\infty}{\longrightarrow} \mu_2$ $(\hat{\mu}_2^{rimp} \overset{p}{\longrightarrow} \mu_2)$ we need these two requirements to hold simultaneously:

- $E(Y_2 \mid y_1, R = 1) = E(Y_2 \mid y_1)$ (implied by MAR)
- $E(Y_2 \mid y_1)$ is correctly specified, i.e., there really exist β_0^* and β_1^* such that $E(Y_2 \mid y_1) = \beta_0^* + \beta_1^* y_1$

However, even if these two conditions hold, single imputation leads to underestimation of variances, as seen with mean imputation

Example of Regression Imputation in Davidian and Tsiatis

Davidian and Tsiatis show that for $\hat{\mu}_2^{rimp} \overset{n\to\infty}{\longrightarrow} \mu_2$ $(\hat{\mu}_2^{rimp} \overset{p}{\longrightarrow} \mu_2)$ we need these two requirements to hold simultaneously:

- $E(Y_2 \mid y_1, R = 1) = E(Y_2 \mid y_1)$ (implied by MAR)
- $E(Y_2 \mid y_1)$ is correctly specified, i.e., there really exist β_0^* and β_1^* such that $E(Y_2 \mid y_1) = \beta_0^* + \beta_1^* y_1$

However, even if these two conditions hold, single imputation leads to underestimation of variances, as seen with mean imputation

Outline

Hot-Deck Imputation

- Replace missing values of a non-respondent (called the recipient) with observed values from a respondent (the donor)
- Recipient and donor need to be similar with respect to variables observed by both cases
  - Donor can be selected randomly from a pool of potential donors
  - Single donor can be identified, e.g. "nearest neighbour" based on some metric
- Andridge & Little (2010, Int. Stat. Rev.) reviewed this approach and concluded that
  - General patterns of missingness are difficult to deal with ("swiss cheese pattern")
  - Lack of theory to support this method
  - Lack of comparisons with other methods
  - Uncertainty from imputation is not taken into account (underestimation of variances)

Hot-Deck Imputation

- Replace missing values of a non-respondent (called the recipient) with observed values from a respondent (the donor)
- Recipient and donor need to be similar with respect to variables observed by both cases
  - Donor can be selected randomly from a pool of potential donors
  - Single donor can be identified, e.g. "nearest neighbour" based on some metric
- Andridge & Little (2010, Int. Stat. Rev.) reviewed this approach and concluded that
  - General patterns of missingness are difficult to deal with ("swiss cheese pattern")
  - Lack of theory to support this method
  - Lack of comparisons with other methods
  - Uncertainty from imputation is not taken into account (underestimation of variances)

Hot-Deck Imputation

- Replace missing values of a non-respondent (called the recipient) with observed values from a respondent (the donor)
- Recipient and donor need to be similar with respect to variables observed by both cases
  - Donor can be selected randomly from a pool of potential donors
  - Single donor can be identified, e.g. "nearest neighbour" based on some metric
- Andridge & Little (2010, Int. Stat. Rev.) reviewed this approach and concluded that
  - General patterns of missingness are difficult to deal with ("swiss cheese pattern")
  - Lack of theory to support this method
  - Lack of comparisons with other methods
  - Uncertainty from imputation is not taken into account (underestimation of variances)

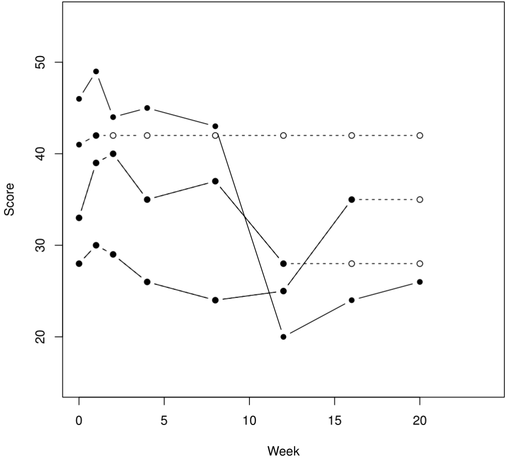
Outline

Last Observation Carried Forward

- Common in settings where a variable is measured repeatedly over time and there is dropout
- If there is droput at time j, we don't observe $Z_j, Z_{j+1}, \ldots, Z_T$
- LOCF: replace all of $Z_j, Z_{j+1}, \ldots, Z_T$ with Z_{j-1}

Last Observation Carried Forward

Example from Davidian and Tsiatis:



Solid lines: observed data. Dashed lines: extrapolated data with LOCF.

Last Observation Carried Forward

Attempts to justify LOCF

- Interest in the last observed outcome measure (reasonable in some context??)
- Under some assumptions, will lead to conservative analysis
  - Say we have a clinical trial, outcome under treatment is expected to improve over time
  - If treatment is found to be superior even with LOCF, then true effect should be even larger
  - Relies on assumption of monotonic improvement over time!

Last Observation Carried Forward

Attempts to justify LOCF

- Interest in the last observed outcome measure (reasonable in some context??)
- Under some assumptions, will lead to conservative analysis
  - Say we have a clinical trial, outcome under treatment is expected to improve over time
  - If treatment is found to be superior even with LOCF, then true effect should be even larger
  - Relies on assumption of monotonic improvement over time!

Example of LOCF in Davidian and Tsiatis

Study participants' characteristic to be measured at T times

- Y_j: measurement taken at time t_j
- D: participant dropout time
- Interest: $\mu_T = E(Y_T)$
- The LOCF estimator of the mean is

$$\hat{\mu}_T^{LOCF} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{T} I(D_i = j+1) Y_{ij}$$

- The expected value of the LOCF estimator of the mean is

$$E(\hat{\mu}_T^{LOCF}) = \mu_T - \sum_{j=1}^{T-1} E[I(D = j+1)(Y_T - Y_j)],$$

so $\hat{\mu}_T^{LOCF}$ is biased, in general

Example of LOCF in Davidian and Tsiatis

Study participants' characteristic to be measured at T times

- Y_j: measurement taken at time t_j
- D: participant dropout time
- Interest: $\mu_T = E(Y_T)$
- The LOCF estimator of the mean is

$$\hat{\mu}_T^{LOCF} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{T} I(D_i = j+1) Y_{ij}$$

- The expected value of the LOCF estimator of the mean is

$$E(\hat{\mu}_T^{LOCF}) = \mu_T - \sum_{j=1}^{T-1} E[I(D = j+1)(Y_T - Y_j)],$$

so $\hat{\mu}_T^{LOCF}$ is biased, in general

Example of LOCF in Davidian and Tsiatis

Study participants' characteristic to be measured at T times

- Y_j: measurement taken at time t_j
- D: participant dropout time
- Interest: $\mu_T = E(Y_T)$
- The LOCF estimator of the mean is

$$\hat{\mu}_T^{LOCF} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{T} I(D_i = j+1)Y_{ij}$$

- The expected value of the LOCF estimator of the mean is

$$E(\hat{\mu}_T^{LOCF}) = \mu_T - \sum_{j=1}^{T-1} E[I(D = j+1)(Y_T - Y_j)],$$

so $\hat{\mu}_T^{LOCF}$ is biased, in general

Example of LOCF in Davidian and Tsiatis

Study participants' characteristic to be measured at T times

- Y_j: measurement taken at time t_j
- D: participant dropout time
- Interest: $\mu_T = E(Y_T)$
- The LOCF estimator of the mean is

$$\hat{\mu}_T^{LOCF} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{T} I(D_i = j+1) Y_{ij}$$

- The expected value of the LOCF estimator of the mean is

$$E(\hat{\mu}_T^{LOCF}) = \mu_T - \sum_{j=1}^{T-1} E[I(D = j+1)(Y_T - Y_j)],$$

so $\hat{\mu}_T^{LOCF}$ is biased, in general

Outline

Summary

Main take-aways from today's lecture:

- Complete-case analyses are wasteful. Also, potentially invalid unless MCAR
- Available-case analyses make a better use of the available data but still requires MCAR (weaker assumptions possibly depend on model/quantity being used/estimated)
- Imputation methods might be valid for some quantities under MCAR but variances are underestimated $\Longrightarrow$ overconfidence in your results!

Next lecture:

- R session 1: imputation methods, some simulation studies
- Bring your laptops!

Summary

Main take-aways from today's lecture:

- Complete-case analyses are wasteful. Also, potentially invalid unless MCAR
- Available-case analyses make a better use of the available data but still requires MCAR (weaker assumptions possibly depend on model/quantity being used/estimated)
- Imputation methods might be valid for some quantities under MCAR but variances are underestimated $\Longrightarrow$ overconfidence in your results!

Next lecture:

- R session 1: imputation methods, some simulation studies
- Bring your laptops!