# Notes: Statistical Analysis with Missing Data – Ch3 Complete Case Analysis and Weighting Methods

Yingbo Li

09/08/2020

# Table of Contents

# Complete-case (CC) analysis

- Complete-case (CC) analysis: use only data points (units) where all variables are observed
- Loss of information in CC analysis:
  - Loss of precision (larger variance)
  - Bias, when the missingness mechanism is not MCAR. In this case, the complete units are not a random sample of the population
- In this notes, I will focus on the bias issue
  - Adjusting for the CC analysis bias using weights
  - This idea is closed related to weighting in randomization inference for finite population surveys

# Notations

- Population size $N$, sample size $n$
- Number of variables (items): $K$
- Data: $Y = (y_{ij})$, where $i = 1, \ldots, N$ and $j = 1, \ldots, K$
- Design information (about sampling or missingness): $Z$
- Sample indicator: $I = (I_1, \ldots, I_N)'$; for unit $i$,

$$I_i = \mathbf{1}_{\{\text{unit } i \text{ included in the sample}\}}$$

- Sample selection processes can be characterized by a distribution for $I$ given $Y$ and $Z$.

# Probability sampling

- Properties of probability sampling
  1. Unconfounded: selection doesn't depend on $Y$, i.e.,

$$f(I \mid Y, Z) = f(I \mid Z)$$

  2. Every unit has a positive (known) probability of selection

$$\pi_i = P(I_i = 1 \mid Z) > 0, \quad \text{for all } i$$

- In equal probability sample design, $\pi_i$ is the same for all $i$

## Stratified random sampling

- $Z$ is a variable defining strata. Suppose Stratum $Z = j$ has $N_j$ units in total, for $j = 1, \ldots, J$
- In Stratum $j$, stratified random sampling takes a simple random sample of $n_j$ units
- The distribution of $I$ under stratified random sampling is

$$f(I \mid Z) = \prod_{j=1}^{J} \binom{N_j}{n_j}^{-1}$$

## Example: estimating population mean $\bar{Y}$

- An unbiased estimate is the stratified sample mean

$$\bar{y}_{\text{st}} = \frac{\sum_{j=1}^{J} N_j \bar{y}_j}{N}$$

where $\bar{y}_j$ is the sample mean in stratum $j$

- Sampling variance approximation

$$v(\bar{y}_{st}) \approx \frac{1}{N^2} \sum_{j=1}^{J} N_j^2 \left( \frac{1}{n_j} - \frac{1}{N_j} \right) s_j^2$$

where $s_j$ is the sample variance of $Y$ in stratum $j$

- A large sample 95% confidence interval for $\bar{Y}$ is

$$\bar{y}_{\text{st}} \pm 1.96 \sqrt{v(\bar{y}_{st})}$$

## Weighting methods

- Main idea: A unit selected with probability $\pi_i$ is "representing" $\pi_i^{-1}$ units in the population, hence should be given weights $\pi_i^{-1}$.
- For example, in stratified random sample
  - A selected unit $i$ in stratum $j$ represents $N_j/n_j$ population units
  - Thus by Horvitz-Thompson estimate, the population mean can be estimated by the weighted sum

$$\bar{y}_w = \frac{1}{n}\sum_{i=1}^{n} w_i y_i, \quad \pi_i = \frac{n_j}{N_j}, \quad w_i = n \cdot \frac{\pi_i^{-1}}{\sum_k \pi_k^{-1}}$$

  - It is not hard to show that

$$\bar{y}_w = \bar{y}_{\text{st}}$$

# Weighting with nonresponses

- If the probability of selecting unit $i$ is $\pi_i$, and the probability of response for unit $i$ is $\phi_i$, then

$$P(\text{unit } i \text{ is observed}) = \pi_i \phi_i$$

- Suppose there are $r$ units observed (respondents). Then the weighted estimate for $\bar{Y}$ is

$$\bar{y}_w = \frac{1}{r}\sum_{i=1}^{r} w_i y_i, \quad w_i = r \cdot \frac{(\pi_i \phi_i)^{-1}}{\sum_k (\pi_k \phi_k)^{-1}}$$

- Usually $\phi_i$ is unknown and thus needs to be estimated

# Weighting class estimator

- Weighting class adjustments are used primarily to handle unit nonresponse
- Suppose we partition the sample into $J$ "weighting classes". In the weighting class $C = j$:
  - $n_j$: the sample size
  - $r_j$: number of observed samples
  - A simple estimator for $\phi_j$ is $\hat{\phi}_j = \frac{r_j}{n_j}$
- For equal probability designs, where $\pi_i$ is constant, the weighting class estimator is
$$\bar{y}_{\text{wc}} = \frac{1}{n} \sum_{j=1}^{J} n_j \bar{y}_{j\text{R}}$$
  where $\bar{y}_{j\text{R}}$ is the respondent mean in class $j$
- The estimate is unbiased under the following form of MAR assumption (Quasirandomization): data are MCAR within weighting class $j$

# More about weighting class adjustments

- Pros: handle bias with one set of weights for multivariate $Y$
- Cons: weighting is inefficient and can increase in sampling variance, if $Y$ is weakly related to the weighting class variable $C$
- How to choose weighting class adjustments: weighting is only effective for outcomes ($Y$) that are associated with the adjustment cell variable ($C$). See the right column in the table below.

**Table 3.1** Example 3.6: effect of weighting adjustments on bias and sampling variance of a mean, by strength of association of the adjustment cell variables with nonresponse and outcome

| | Association with outcome | |
|---|---|---|
| **Association with nonresponse** | **Low (L)** | **High (H)** |
| Low (L) | Bias: — | Bias: — |
| | Var: — | Var: ↓ |
| High (H) | Bias: — | Bias: ↓ |
| | Var: ↑ | Var: ↓ |

# Propensity weighting

- The theory of propensity scores provides a prescription for choosing the coarsest reduction of $X$ to a weighting class variable $C$ so that quasirandomization is roughly satisfied
- Let $X$ denote the variables observed for both respondents and nonrespondents
- Suppose data are MAR, with $\phi$ being unknown parameters about missing mechanism

$$P(M \mid X, Y, \phi) = P(M \mid X, \phi)$$

Then quasirandomization is satisfied when $C$ is chosen to be $X$

# Response propensity stratification

- Define response propensity for unit $i$ as

$$\rho(x_i, \phi) = P\left(m_i = 0 \mid \rho(x_i, \phi), \phi\right)$$

  i.e., respondents are a random subsample within strata defined by the propensity score $\rho(X, \phi)$

- Usually $\phi$ is unknown. So **a practical procedure** is

  (i) Estimate $\hat{\phi}$ from a binary regression of $M$ on $X$, based on respondent and nonrespondent data

  (ii) Let $C$ be a grouped variable by coarsening $\rho\left(X, \hat{\phi}\right)$ into 5 or 10 values

- Thus, within the same adjustment class, all respondents and nonrespondents have the same value of the grouped propensity score

# An alternative procedure: propensity weighting

- An alternative procedure is to weight respondents $i$ directly by the inverse propensity score $\rho\left(X, \hat{\phi}\right)^{-1}$
- This method removes nonresponse bias
- But it may yield estimates with extremely high sampling variance because respondents with very low estimated response propensities receive large nonresponse weights
- Also, weighting directly by inverse propensities place may reliance on correct model specification of the regression of $M$ on $X$

## Example: inverse probability weighted generalized estimating equations (GEE)

- Let $x_i$ be covariates of GEE, and $z_i$ be a fully observed vector that can predict missing mechanism
- If $P(m_i = 1 \mid x_i, y_i, z_i, \phi) = P(m_i = 1 \mid x_i, \phi)$, then the unweighted completed case GEE is unbiased

$$\sum_{i=1}^{r} D_i(x_i, \beta)\,[y_i - g(x_i, \beta)] = 0$$

- If $P(m_i = 1 \mid x_i, y_i, z_i, \phi) = P(m_i = 1 \mid x_i, z_i, \phi)$, then the inverse probability weighted GEE is unbiased

$$\sum_{i=1}^{r} w_i(\hat{\alpha}) D_i(x_i, \beta)\,[y_i - g(x_i, \beta)] = 0, \quad w_i(\hat{\alpha}) = \frac{1}{p(x_i, z_i \mid \hat{\alpha})}$$

where $p(x_i, z_i \mid \hat{\alpha})$ is the probability of being a complete unit, based on logistic regression of $m_i$ on $x_i, z_i$

# Poststratification

- The weighting class estimator

$$\bar{y}_{\text{wc}} = \frac{1}{n} \sum_{j=1}^{J} n_j \bar{y}_{j\text{R}}$$

  uses the sample proportion $n_j/n$ to estimate the population proportion $N_j/N$.

- If from an external resource (e.g., census or a large survey), we know the population proportion of weighting classes, then we can use the post stratified mean to estimate $\bar{Y}$:

$$\bar{y}_{\text{ps}} = \frac{1}{N} \sum_{j=1}^{J} N_j \bar{y}_{j\text{R}}$$

# Summary of weighting methods

- Weighted CC estimates are often simple to compute, but the appropriate standard errors can be hard to compute (even asymptotically)
- Weighting methods treat weights as fixed and known, but these nonresponse weights are computed from observed data and hence are subject to sampling uncertainty
- Because weighted CC methods discard incomplete units and do not provide an automatic control of sampling variance, they are most useful when
  - Number of covariates is small, and
  - Sample size is large

# Available-case (AC) analysis

- Available-case analysis: for univariate analysis, include all unites where that variable is present
  - Sample changes from variable to variable according to the pattern of missing data
  - This is problematic if not MCAR
  - Under MCAR, AC can be used to estimate mean and variance for a single variable
- Pairwise AC: estimates covariance of $Y_j$ and $Y_k$ based on units $i$ where both $y_{ij}$ and $y_{ik}$ are observed
  - Pairwise covariance estimator:

$$s_{jk}^{(jk)} = \sum_{i \in I_{jk}} \left( y_{ij} - \bar{y}_j^{(jk)} \right) \left( y_{ik} - \bar{y}_k^{(jk)} \right) / \left( n^{(jk)} - 1 \right)$$

where $I_{jk}$ is the set of $n^{(jk)}$ units with both $Y_j$ and $Y_k$ observed

## Problems with pairwise AC estimators on correlation

- Correlation estimator 1:

$$r^*_{jk} = \frac{s^{(jk)}_{jk}}{\sqrt{s^{(j)}_{jj} s^{(k)}_{kk}}}$$

  - Problem: it can lie outside of $(-1, 1)$

- Correlation estimator 2 corrects the previous problem:

$$r^{(jk)}_{jk} = \frac{s^{(jk)}_{jk}}{\sqrt{s^{(jk)}_{jj} s^{(jk)}_{kk}}}$$

- Under MCAR, all these estimators on covariance and correlation are consistent

- However, when $K > 3$, both correlation estimators can yield correlation matrices that are not positive definite!
  - An extreme example: $r_{12} = 1, r_{13} = 1, r_{23} = -1$

## Compare CC and AC methods

- When data is MCAR and correlations are mild, AC methods are more efficient than CC
- When correlations are large, CC methods are usually better

# References


- Little, R. J., & Rubin, D. B. (2019). Statistical Analysis with Missing Data, 3rd Edition. John Wiley & Sons.