Incomplete Data Analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2020/2021

Naive methods

- We will review some naive methods for handling missing data. Although, somewhat ad hoc, these methods have been widely used in practice and so they deserve our attention.
- These methods deal with missing data either by **removing the cases/individuals with incomplete data** or by **filling in the missing values (one single value is used–single imputation)**.
- In this part we will focus in the first class of methods (those disregarding, totally or partially, cases with incomplete data), namely
  - Complete case analysis.
  - Available case analysis.

Naive methods

Bias of an estimator – definition

- We will be mentioning that some of the approaches have the potential to induce bias. It is then worth defining bias before proceeding.

- Suppose that $Y_1, \ldots, Y_n$ are iid random variables, each with pdf/pmf $f_Y(y \mid \theta)$, with θ unknown.

- If $\widehat{\theta} = T(Y_1, \ldots, Y_n)$ is an estimator of θ, then the bias of $\widehat{\theta}$ is the difference between its expectation and the 'true' value, i.e.,

$$\text{bias}(\widehat{\theta}) = E(\widehat{\theta}) - \theta.$$

- An estimator $\widehat{\theta} = T(Y_1, \ldots, Y_n)$ is unbiased if

$$E(\widehat{\theta}) = \theta,$$

  otherwise it is biased.

Naive methods

Bias of an estimator – example

- Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathrm{N}(\theta, 1)$.
- A possible estimator for θ is

$$\widehat{\theta} = T(Y_1, \ldots, Y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

- We have

$$\begin{aligned} E(\widehat{\theta}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(Y_i) \\ &= \theta. \end{aligned}$$

- Thus, $\widehat{\theta}$ is unbiased for θ.

Naive methods

Complete case analysis

- **Complete case analysis** (also known as **listwise deletion**) excludes the data for any case/individual that has one or more missing values.

- The data are treated as if the cases with missing values were not there. In the table below, the second, third, and fourth subjects would be discarded.

Gender	Age	Income	...
M	28	$80,000$	...
F	?	?	...
?	43	?	...
F	?	$100,000$	...

- This method is, possibly, one of the most used in the applied (medical, social, etc) sciences.

Naive methods

Complete case analysis

- ↪ Restricting the analyses to complete cases eliminates the need for specialised software and for advanced missing data handling procedures.
- ↪ But what are the costs associated to such simplicity?
- ↪ If data are MCAR, because the observed data are a random sample from the complete data, complete cases analysis lead to valid inferences.
- ↪ However, standard error will be larger than in the case of no missing data, so that confidence intervals will be wider and power reduced, compared with the no missing data situation.
- ↪ Further, complete case analysis is potentially wasteful. If the number of variables is large, there may be very few complete cases, so that most of the data would be discarded for the sake of a complete analysis.

Naive methods

Complete case analysis

- It is legitimate to ask if there is a threshold on the percentage of missing cases below which a complete case analysis is recommended.

- There are answers for all tastes! At one end of the spectrum we find Enders (2010, p. 39):

  "*In most situations, the disadvantages of listwise deletion far outweigh its advantages.*"

- Graham (2009, p. 554, *Annual Review of Psychology*) covers the middle ground:

  "*I, personally, would still use one of the missing data approaches even with just* 5% *missing cases and I encourage you to get used to doing the same. However, if a researcher choose to stay with a listwise deletion under these special circumstances I believe it would be unreasonable for a critic to argue that it was a bad idea to do so.*"

Naive methods

Complete case analysis

- ↪ Even if the data are ‘only’ MAR, inferences from a complete case analysis can be biased.
- ↪ However, in fairness, complete case analysis is not always bad.
- ↪ In the context of regression analysis, complete case analysis possesses some unique properties that make it attractive in particular settings.
- ↪ In particular, complete case analysis can produce unbiased estimates of regression coefficients under any missing data mechanism, provided that the probability of observing a complete case (an individual for whom both the response variable and the covariates are observed) is a function of the covariates and not of the response variable.
- ↪ We should, nevertheless, keep in mind that this very particular result is one of the very few situations in which complete case analysis is likely to outperform more advanced techniques (e.g., maximum likelihood and multiple imputation) for MAR/MNAR data.

Naive methods

Available case analysis

- **Available case analysis**, also known as **pairwise deletion**, attempts to mitigate the data loss problem of complete case analysis.

- In available case analysis different aspects of a problem are studied with different subsets of the data.

- The prototypical application of available case analysis is the computation of covariance/correlation matrices, where different subsets of cases are used to compute each element in the covariance/correlation matrix.

Naive methods

Available case analysis

- Consider as an example, a simple two variable (X_1, X_2) data matrix with only one variable, X_2, subject to missingness.

- In available case analysis, all cases would be used to estimate the mean and variance of X_1, but only the complete cases would contribute to an estimate of the mean and variance of X_2 and the covariance between X_1 and X_2.

$$\left.\begin{array}{ll} x_{11} & x_{21} \\ x_{12} & x_{22} \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ x_{1m} & x_{2m} \end{array}\right\} m \text{ Complete Cases}$$

$$\left.\begin{array}{ll} x_{1(m+1)} & - \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ x_{1n} & - \end{array}\right\} n - m \text{ Cases with observations on } x_1$$

Figure from Pigott (2001).

Naive methods

Available case analysis

↪ The available case estimates are

$$\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{1i},$$

$$\bar{x}_2 = \frac{1}{m}\sum_{i=1}^{m} x_{2i},$$

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2,$$

$$s_2^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_{2i} - \bar{x}_2)^2,$$

$$s_{1,2} = \frac{1}{m-1}\sum_{i=1}^{m}(x_{1i} - \bar{x}_{1(m)})(x_{2i} - \bar{x}_2),$$

where $\bar{x}_{1(m)}$ is the mean calculated from the m cases.

Naive methods

Available case analysis

- The covariance matrix is then given by

$$\begin{bmatrix} s_1^2 & s_{1,2} \\ s_{1,2} & s_2^2 \end{bmatrix}$$

- Similarly, the correlation between X_1 and X_2, under available case analysis, would be

$$\rho_{1,2} = \frac{s_{1,2}}{s_{1(m)} s_2}.$$

- This method is simple, uses all available information and produces consistent estimates of mean, correlations and covariances under MCAR (Little and Rubin, 2002, p.55).

- Similarly to complete case analysis, available case analysis can also lead to biased estimates if the data are not MCAR.

Naive methods

Available case analysis

- Available case analysis have also a number of unique problems that limits its utility.

- Possibly, the major of such problems is that the resulting covariance or correlation matrix may not be a positive semi definite one which is, however, a requirement for various subsequent analysis processing covariance/correlation matrices (e.g., principal component analysis).

Naive methods

Toy example

↪ Let us consider the following hypothetical dataset with 10 individuals and 3 variables and compute the covariance matrix under the complete case analysis and available case analysis approaches.

Y_1	Y_2	Y_3
26	56	NA
25	NA	158
20	40	NA
NA	49	158
24	NA	164
20	43	134
NA	50	161
NA	48	NA
21	NA	134
25	53	169

Naive methods

Toy example

- ↪ Following a complete case analysis strategy only the cases in row 6 and 10 enter the analysis.

- ↪ All means, variances and covariances are based only on the data from these two individuals.

- ↪ The resulting covariance matrix is given by

$$\Sigma_{\text{CCA}} = \begin{pmatrix} 12.5 & 25 & 87.5 \\ 25 & 50 & 175 \\ 87.5 & 175 & 612.5 \end{pmatrix}.$$

Naive methods

Toy example

- ↪ Following an available case analysis, the variances will be based on data from 7 individuals (which are not the same for the three variables).

- ↪ For the covariances of the first and second variable and the second and third variable, four pairs of values are available. In turn, for the covariance between the first and third variables, five pairs of values are available.

- ↪ The resulting covariance matrix (with entries rounded to two decimal places) is given by

$$\Sigma_{\mathrm{ACA}} = \begin{pmatrix} 6.67 & 24.33 & 37 \\ 24.33 & 30.29 & 62.83 \\ 37 & 62.83 & 201 \end{pmatrix}.$$