Strategies for handling missing data caused by item nonresponse in environmental monitoring programs

Michael Dumelle[a,*], Anthony R. Olsen[a], Amanda Nahlik[a], Karen Blocksom[a], Other[b]

[a] *United States Environmental Protection Agency, 200 SW 35th St, Corvallis, 97330*

[b] *Other,*

Abstract

This is the abstract. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum augue turpis, dictum non malesuada a, volutpat eget velit. Nam placerat turpis purus, eu tristique ex tincidunt et. Mauris sed augue eget turpis ultrices tincidunt. Sed et mi in leo porta egestas. Aliquam non laoreet velit. Nunc quis ex vitae eros aliquet auctor nec ac libero. Duis laoreet sapien eu mi luctus, in bibendum leo molestie. Sed hendrerit diam diam, ac dapibus nisl volutpat vitae. Aliquam bibendum varius libero, eu efficitur justo rutrum at. Sed at tempus elit.

Keywords: keyword1, keyword2

1. Introduction

- Item nonresponse

Rubin (1976) categorizes missing data into three distinct types: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Data are MCAR if the probability of missingness is the same for each observation. Data are MAR if the probability of missingness is the same for each observation belonging to a group defined by the observed data. Data are MNAR if the probability of missingness is related to the data. Suppose a researcher is using a remote sensor to track daily the proportion of lake area covered by algae. Occasionally, the remote sensor malfunctions and data cannot be collected. If the probability that the remote sensor malfunctions is completely random and the same for each day, the lake algae data are MCAR. Suppose that the remote sensor malfunctions more often on cloudy days than sunny days, but the researcher knows whether each day is sunny or cloudy. The lake algae data are MAR if the probability the remote sensor malfunctions is 1) the same for all cloudy days and 2) the same for all sunny days. Now suppose that the remote sensor malfunctions more often on humid days than non-humid days, but the researcher does not have access to humidity status for each day. Because the probability the remote sensor malfunctions is related to humidity status, which cannot be observed, the data are MNAR.

2. Background

There are two types of item nonresponse we consider: structural and non-structural. Item nonresponse is structural when the reason that the data are missing is directly related to physical features at a site.

*Corresponding author

Email address: Dumelle.Michael@epa.gov (Michael Dumelle)

When structural nonresponse occurs, the data are "not missing at random", but we have information about the mechanism that causes the missingness. Item nonresponse is non-structural when the reason the data are missing is not directly related to physical features at a site. The approaches for handling structural vs non-structural item nonresponse vary drastically, so it is important the correctly identify the type of nonresponse and apply appropriate analysis techniques.

Structural item nonresponse

This is a citation (Rubin, 1996)

A straightforward approach to handling missing data is complete case analysis, sometimes called listwise deletion. Complete case analysis involves removing all observations from the data that have missingness in at least one variable. There are several benefits to complete case analysis. Complete case analysis is intuitive and computationally simple. When the data are MCAR, complete case analysis yields unbiased estimates of means (Little and Rubin, 2019). Unfortunately, there are also some drawbacks to complete case analysis. Complete case analysis is inefficient and can lead to a significant loss of valuable information. Consider a scenario where 100 variables are measured for 100 observations. Suppose that each observation is missing data for only one variable, and that each observation is missing data for a different variable. Though 99% of the data are observed, complete case analysis would throw out every observation, resulting in no usable data. Additionally, complete case analysis yields biased mean estimates when the data are not MCAR (Little, 1992; Schafer and Graham, 2002; Rubin, 2004; Donders et al., 2006; White and Carlin, 2010; Little and Rubin, 2019). In spite of these drawbacks, complete case analysis is commonly used.

In single imputation, missing data are first replaced via some imputation method. Then this single set of data (which no longer has missingness) is analyzed as if it were actually observed without missingness. Unconditional mean imputation involves replacing missing values of a variable with the mean of the variable in the observed data. Conditional mean imputation involves replacing missing values of a variable with the mean of the variable in the observed data conditional on some auxiliary variable. Consider a field experiment measuring nitrogen content in soil after exposure to a control or treatment group. Unconditional mean imputation replaces missing data by the average nitrogen content, regardless of whether the missingness occurred in the control or treatment group. Conditional mean imputation replaces missing data by the average nitrogen content within each group (treatment or control).

3. Applications to National Aquatic Resource Survey Data

4. Discussion

5. Quarto Examples

5.1. Equations

Here is an equation:

$$f_X(x) = \left(\frac{\alpha}{\beta}\right)\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}}; \alpha, \beta, x > 0. \tag{1}$$

It is Equation 1.

Inline equations work as well: $\sum_{i=2}^{\infty}\{\alpha_i^{\beta}\}$

5.2. Figures and tables

Figure 1 is generated using an R chunk.

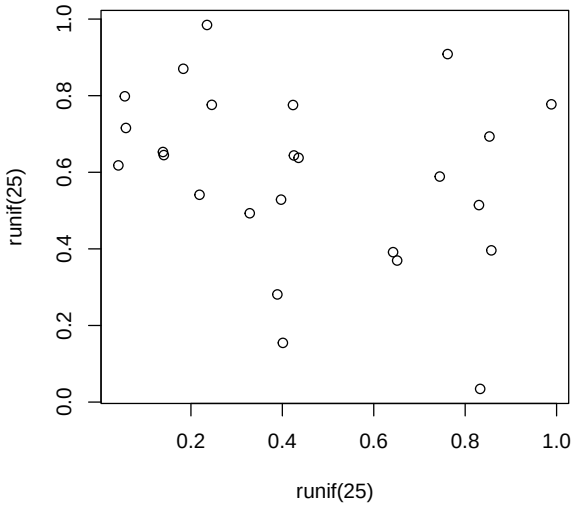

Figure 1: A meaningless scatterplot

5.3. Tables coming from R

Tables can also be generated using R chunks, as shown in Table 1 example.

```
knitr::kable(head(mtcars)[,1:4])
```

Table 1: Caption centered above table

	mpg	cyl	disp	hp
Mazda RX4	21.0	6	160	110
Mazda RX4 Wag	21.0	6	160	110
Datsun 710	22.8	4	108	93
Hornet 4 Drive	21.4	6	258	110
Hornet Sportabout	18.7	8	360	175
Valiant	18.1	6	225	105

References

Donders, A.R.T., Van Der Heijden, G.J., Stijnen, T., Moons, K.G., 2006. A gentle introduction to imputation of missing values. Journal of clinical epidemiology 59, 1087–1091.

Little, R.J., 1992. Regression with missing x's: a review. Journal of the American statistical association 87, 1227–1237.

Little, R.J., Rubin, D.B., 2019. Statistical analysis with missing data. volume 793. John Wiley & Sons.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–592.

Rubin, D.B., 1996. Multiple imputation after 18+ years. Journal of the American statistical Association 91, 473–489.

Rubin, D.B., 2004. Multiple imputation for nonresponse in surveys. volume 81. John Wiley & Sons.

Schafer, J.L., Graham, J.W., 2002. Missing data: our view of the state of the art. Psychological methods 7, 147.

White, I.R., Carlin, J.B., 2010. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statistics in medicine 29, 2920–2931.