Missing data 4. Missing data in the outcomes versus in the covariates

Tra My Pham, Nikolaos Pandis, Ian R White

Missing data can occur in more than one variable, such as in the outcome, one or more covariates, or very often both. In this article, we explain why the distinction between missing data in the outcomes versus in the covariates is important for choosing a way to analyse the data. In particular, we discuss its implication on choosing whether to use the simplest analysis option, a complete case analysis (described below).

In randomised controlled trials, several situations can give rise to missing data in the outcome.[1] For many different reasons, individuals may stop participating in a trial and withdraw consent for further data collection, after which point they no longer provide data on the outcome variable of interest. Individuals may also fail to attend some follow-up visits; hence, for such visits their outcome is not measured. In addition to missing data in the outcome, we might fail to collect all the required data for some individuals during their visit at baseline, resulting in some baseline variables being partially observed (i.e. missing values in the covariates). Likewise, in observational studies, missing data can occur in both the outcome and covariates, with missing values in the covariates being a very prevalent issue.

A complete case analysis (CCA) is a default method for handling missing values in common statistical software packages and is easy to perform. In a CCA, individuals with missing values in any of the variables considered in the analysis are excluded, and the analysis is performed on individuals with complete data in all variables. In settings where the analysis consists of fitting a regression model, a CCA is valid when any of the following is true:

- The data are missing completely at random (MCAR);
- The outcome is missing at random (MAR), conditional on the covariates included in the model, and the covariates are fully observed;
- The covariates are partially observed, but conditional on the values of the covariates, missingness in the covariates does not depend on the values of the outcome.[2–4]

The first two scenarios where a CCA is valid are widely known, while the third case refers to a different sort of assumption. However, there might be situations where the third case is more plausible. For example, in cohort studies where individuals are followed up over time, it might not be plausible to assume that missingness in a covariate measured at baseline is caused by the outcome that will only be measured in the future. Instead, it might seem more likely that missingness in the covariate is either caused by the values of the covariate, or by other variables also measured at baseline.

Whether a CCA is valid is not the only consideration. Even in situations when a CCA is valid, estimates obtained from a CCA suffer from a loss of precision, and other methods may be better. This is especially relevant when missing data occur in the confounders.

We will illustrate this with an example created using data from a cohort study conducted to assess whether gingival recession is more likely in individuals who had orthodontic treatment compared to those without orthodontic treatment.[5] The outcome variable is *recession_sum_33b* which is the number of teeth with buccal gingival recession among the lower six front teeth; this variable takes

values 0 to 6. The exposure variable is *group* (orthodontic treatment versus no orthodontic treatment). We also have data on the individual's gender and age at the end of treatment. Since gender is a potential confounder, we want to adjust for gender when estimating the association between orthodontic treatment and recession. Our analysis model is therefore a linear regression of *recession_sum_33b* on *group* and *gender*. Let us suppose that our full data set contains 190 individuals, with fully observed data on all three variables included in the analysis. Then, we artificially make 50% of values in *gender* missing under the MCAR assumption, resulting in a CCA data set of 95 individuals with fully observed data on recession, treatment, and gender. The full-data and CCA results are presented in Table 1.

Table 1. Results from a linear regression of *recession_sum_33b* on *group* and adjusting for *gender*, fitted to the full data (left panel) and complete cases (right panel)

	Full data (N=190)			Complete case analysis (N=95)		
	Coeff-icient	Standard error	95% confidence interval	Coeff-icient	Standard error	95% confidence interval
Group	-0.91	0.25	(-1.40, -0.41)	-0.93	0.38	(-1.67, -0.18)
Gender	-0.47	0.26	(-0.99, 0.05)	-0.65	0.39	(-1.42, 0.13)

Since *gender* is MCAR, there is no bias from a CCA. Differences in the point estimates for both treatment group and gender between CCA and the full data are small, reflecting the smaller amount of information used in the CCA. However, CCA produces larger standard errors, and consequently wider 95% confidence intervals (less precision) compared with those obtained from the full-data analysis. This applies even for the coefficient of the variable *group* which has complete data. This is therefore an analysis for which multiple imputation (MI) would be valuable because it will improve the precision of the estimates.

When CCA is not desirable, it does not follow that MI is the only alternative. One such setting is in randomised controlled trials, where adjustment for baseline covariates can increase power to detect a treatment effect. When baseline covariates are partially observed, simple methods such as replacing the missing values with the mean of the observed values (i.e. mean imputation), or including as a covariate an indicator variable for missingness (i.e. missing indicator), might be appropriate.[6] In contrast, these simple methods are almost always never valid for observational studies. A second setting where MI is not the only valid approach is when missing values are only in the outcome. Missing outcome data may be handled by a suitable analysis of the available data. For an outcome measured at one time, a CCA is appropriate;[7] for a repeatedly measured outcome, a mixed model could be used for analysis as it intrinsically accounts for the missing values under a MAR assumption.[8] However, when missing data occur in several variables and the MAR assumption seems plausible, one should consider using MI to handle the missing values. The next article will explain the principles of MI.

References

1. National Research Council. *The prevention and treatment of missing data in clinical trials*, www.nap.edu (2010).

2. Carpenter JR, Smuk M. Missing data: a statistical framework for practice. *Biometrical J* 2021;

1–33.

3. Bartlett JW, Harel O, Carpenter JR. Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. *Am J Epidemiol* 2015; 182: 730–736.

4. Bartlett JW, Carpenter JR, Tilling K, et al. Improving upon the efficiency of complete case analysis when covariates are MNAR. *Biostatistics* 2014; 15: 719–730.

5. Gebistorf M, Mijuskovic M, Pandis N, et al. Gingival recession in orthodontic patients 10 to 15 years posttreatment: A retrospective cohort study. *Am J Orthod Dentofac Orthop* 2018; 153: 645–655.

6. White IR, Thompson SG. Adjusting for partially missing baseline measurements in randomized trials. *Stat Med* 2005; 24: 993–1007.

7. Groenwold RH, Donders AR, Roes KC, et al. Dealing with missing outcome data in randomized trials and observational studies. *Am J Epidemiol* 2011; 175: 210--2017.

8. White IR, Moodie E, Thompson SG, et al. A modelling strategy for the analysis of clinical trials with partly missing longitudinal data. *Int J Methods Psychiatr Res* 2003; 12: 139–150.