Methods for Decision Analysis in Missing Data Scenarios

Adway Wadekar

Advisor: Prof. Jerry Reiter

October 1, 2024

1 Introduction

Missingness is a ubiquitous attribute of most datasets arising in public health, policy, and more broadly, in the social and natural sciences [LR19; Rub76]. For example, in a data set about test performance in a school district, data for some students may not be available due to some students being sick that day. Or, in the case of a clinical trial, data for some patients' follow-up lab results may not be available because they may not have the resources to get to a hospital each week. Yet another source of missingness could be, in the case of the school district, if students who feel that they may perform poorly choose to stay home that day.

All three of these basic examples of missingness correspond to three different *types*, which Rubin [LR19] formulates as missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). Following Rubin's notational setup, let $Y = (Y_{\text{obs}}, Y_{\text{miss}})$ be a vector of individuals in a sample, which contains the values that an analyst observes and the values that the analyst does not observe. Let $R = (r_1, \ldots, r_n)$ be a corresponding vector for those individuals, where $r_i = 1$ if the individual has a missing value and $r_i = 0$. Let θ be the parameters associated with Y and ϕ be the (distinct) parameters associated with R.

Definition 1.1 (Missing completely at random)**.**

$$f(R \mid Y, \theta, \phi) = f(R \mid \phi) \tag{1}$$

In data that are MCAR, the missingness pattern is not explained by any of the actual values or the parameters of their distribution.

Definition 1.2 (Missing at random)**.**

$$f(R \mid Y, \theta, \phi) = f(R \mid Y_{obs}, \phi) \tag{2}$$

In data that are MAR, the missingness pattern is fully explained by the observed values and other variables.

Definition 1.3 (Not missing at random)**.**

$$f(R \mid Y, \theta, \phi) = f(R \mid Y_{obs}, Y_{miss}, \phi) \tag{3}$$

In data that are NMAR, the missingness pattern is a function of both observed and unobserved data. In general, it is extremely difficult to tell whether the type of missingness of data. MCAR is the rarest type of missingness, and it is certainly possible that data in many common settings are NMAR. A common approach is to treat the data as MAR if there is enough available, and draw inferences from multiply imputed data sets according to some form of Rubin's rules [RR07]. However, in recent years, there has been a renewed focus in evaluating whether this assumption is one that leads to valid conclusions. In particular, in the case of MAR vs. MNAR, using multiple imputation, a model-based imputation approach that relies only on observed data can lead to biased conclusions when data are MAR and in fact are MAR. [HZ12] consider a related problem, where they use posterior predictive checks to to quantify how inadequate an imputation model may be. When an imputation model produces inadequate results, there could be two sources for the inadequacy: (1) the model itself is inadequate or (2) the data are not missing at random.

[Dua+24] develop one of the first testing approaches for determining whether data are MAR or NMAR, but in some sense, determining with high probability whether the missingness pattern is due to a MAR or MNAR mechanism doesn't essentially change the analyst's approach to imputation and subsequent inference, and the missingness itself may not fit into just one category. Moreover, regardless of the imputation strategy used, the results will be biased for any one data set, even though in a perfect world, an imputation strategy may yield an asymptotically unbiased result. Therefore, the goal of this thesis is to study how *decision analyses* [Ber13] can be affected by missingness type and imputation strategy.

In other words, for a given outcome with missingness, how much does the missingness pattern need to be dependent on the outcome itself for the utility for a particular decision to change? This is similar in spirit to tipping point analysis [LR14] and to e-value analysis [VD17; VDM19], where the latter quntifies how much effect an unmeasured confounder must have to explain away the effect of another variable associated with the outcome. We seek to measure how much "missingness" must be correlated with the outcome of interest in order for the imputation method to change the outcome.

2 Outline of approach

We begin by conducting initial simulations with a simple example of a political poll. In this simulation, the outcome $Y = (Y_{\text{obs}}, Y_{\text{miss}})$ consists of data on individuals voting, where the voter status of any observed individual has some fixed effect and for missing individual, there is another additive fixed effect. We will perform multiple imputation testing a grid of additive effects, using Rubin's rules to infer the percentage voting for a given candidate, ultimately comparing this to standard multiple imputation which assumes no extra effect on the outcome for missing values.

Our ultimate goal is to repeat this sort of analysis for a variety of utility functions and scenarios: examples include political polling, medical decision analyses and government policies regarding the spread of infectious diseases. The initial direction proposed is necessarily simplistic in that it there is only one outcome and no covariates; as the project proceeds, our goal is to develop a tool to do this sort of analysis in real scenarios. To test our approaches, we will use public use data which contain missingness (of which there are plenty of sources available). As we are not using human subjects, we consider our ethical implications to be minimal.

References

[Ber13] James Berger. *Statistical decision theory: foundations, concepts, and methods.* Springer Science & Business Media, 2013.

[Dua+24] Rui Duan et al. “Testing the missing at random assumption in generalized linear models in the presence of instrumental variables”. In: *Scandinavian Journal of Statistics* 51.1 (2024), pp. 334–354.

[HZ12] Yulei He and Alan M Zaslavsky. “Diagnosing imputation models by applying target analyses to posterior replicates of completed data”. In: *Statistics in medicine* 31.1 (2012), pp. 1–18.

[LR14] Victoria Liublinska and Donald B Rubin. “Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial”. In: *Statistics in medicine* 33.24 (2014), pp. 4170–4185.

[LR19] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data.* Vol. 793. John Wiley & Sons, 2019.

[RR07] Jerome P Reiter and Trivellore E Raghunathan. “The multiple adaptations of multiple imputation”. In: *Journal of the American Statistical Association* 102.480 (2007), pp. 1462–1471.

[Rub76] Donald B Rubin. “Inference and missing data”. In: *Biometrika* 63.3 (1976), pp. 581–592.

[VD17] Tyler J VanderWeele and Peng Ding. “Sensitivity analysis in observational research: introducing the E-value”. In: *Annals of internal medicine* 167.4 (2017), pp. 268–274.

[VDM19] Tyler J VanderWeele, Peng Ding, and Maya Mathur. “Technical considerations in the use of the E-value”. In: *Journal of Causal Inference* 7.2 (2019), p. 20180007.