NA-Missing Value Processes Notes

Paul A. Hodgetts

23 October 2020

Contents

1 Notes

1.1 A Criticsm of the word *"missing"* as it relates to values

I have italicized and put the word missing in quotations as I find the word to be a misnomer in this situation. The value is only missing in the sense that a person is prevented from performing the analysis they wish to perform. There is still value contained within the variable, and can indeed tell someone quite a bit about the world. It is not missing in the way that a remote or keys can be missing. It is a recorded yet unobserved value.

1.2 Types of *"missing"* values: "Another example of the law that statisticians should not be allowed to make terminology." (McElreath, 2019b, 38:17)

MCAR (Missing Completely At Random): A variable is *"missing"* completely purely by chance. It is unconditionally independent of the *"missingness"* mechanism (Gelman et al., 2020; McElreath, 2019b; Wu & Thompson, 2020). Is constant for all variables or i in second equation below (King, 2020; Wu & Thompson, 2019). In such situations you do not *need* to impute, but imputation does add precision (McElreath, 2019b). *MCAR* does allow for *complete case analysis.*

$$P(M|D) = P(M)$$

$$\tau_i = \tau$$

MAR (Missing At Random): Implies that the *"missingness"* of a variable is dependent upon or related to some of the observed data (Gelman et al., 2020; McElreath, 2019b; Wu & Thompson, 2020). Is dependent upon only M or x_i in the equations below (King, 2020; Wu & Thompson, 2020). This situation generates a pattern, where other features or variables predict *"missing"* values and are associated causally with the *"missingness"* (McElreath, 2019b; Wu & Thompson, 2020). Deletion of *"missing"* values in the case of MAR will cause biased estimates, so will need to do imputation to de-bias estimates (McElreath, 2019a; McElreath, 2019b).

$$P(M|D) \equiv P(M|D_{obs}, D_{mis}) = P(M|D)$$

$$\tau_i = \tau(x_i)$$

MNAR (Missing Not At Random): MNAR indicates that the variable itself influences the *"missingness"* (Gelman et al., 2020; Wu & Thompson, 2020). In MCAR, the *"missingness"* is dependent upon both y_i and x_i in the equations below (Wu & Thompson, 2020). The *"missingness"* mechanism is not ignorable (McElreath, 2019b). This situation may also arise from unobserved variables (or predictors) (Gelman et al., 2020; McElreath, 2019b). In this case, the *"missingness"* is dependent upon information that has not been recorded at all. Gelman et al., (2020) break MNAR into two situations. Namely, missingness that depends on unobserved predictors and missingness that depends on the missing value itself.

$$\tau_i = \tau(y_i, x_i)$$

1.3 Reasons for *"missing"* values

A plethora of reasons exist as to why a value may appear *"missing"*. These include, equipment errors, entry errors (skipped entry as opposed to incorrect value), software errors, mis-match in different datasets, and nonresponses to name a few. Some of the reasons are more technical in their nature, namely equipment error and software error, while others are subject based, namely nonresponse. Although, they can similar results.

Focusing on nonresponse type, Wu and Thompson (2020) present two types of nonresponse for survey data: unit nonresponse and item nonresponse. Unit nonresponse refers to situations in which the unit is selected by the sampling procedure, but nothing is measured (Wu & Thompson, 2020). For instance, unit nonresponse would be a complete survey missing. Item nonresponse refers to situations in which the value is *"missing"* or unknown for a particular variable (Wu & Thompson, 2020). This can be due to an issue in the data collection process, or the refusal of the respondent.

Mis-match between data can create missing values when two, or more, datasets are joined and values exist in one that do not exist in another leading to a return of the demarcation of a *"missing"* value. For instance, in R, these values would be returned as `NA`. There are various reasons these results can occur. For instance, a mis-match in value input could lead to the creation of a *"missing"* value (i.e. matching two datasets of city names and a spelling error in one dataset creates a *"missing"* value in the joined dataset). In this case, resolving the *"missing"* value is a different process than would be performed for the nonresponse types.

1.4 Resolving *"missing"* values

An issue with *"missing"* values is that they can prevent an analysis from proceeding or being accurate. So, what are the means of proceeding with *"missing"* values?

One common suggestion to resolving *"missing"* values (and the one I have been most told to do) is the complete-case analysis. In this situation, all cases containing *"missing"* values are dropped from the dataset and excluded from the analysis (Gelman et al., 2020; McElreath, 2019b). This should be done in only very specific circumstances (namely when *"missing"* values are MCAR) as it can lead to a few problems (McElreath 2019b). For instance, complete-case analysis can mean that a lot of information is discarded leaving very few cases for analysis and biased estimates (Gelman et al., 2020; McElreath, 2019a). Additionally, that if the *"missing"* values differ systematically from the observed cases, the results will be a biased estimate (Gelman et al., 2020).

A second alternative I have heard mentioned, and heard recommended from time-to-time, is to replace missing values using the mean of the values within that column. This should not be done, and should not be offered as a solution. I only write about it here because I have seen and head it offered as such. When replacing a *"missing"* value with the mean, you are claiming to know that value (McElreath, 2019b). The model does not know that this value does not exist and will not register an imputation error associated with the mean value (McElreath, 2019b). This can distort the distribution for the variable leading to underestimations of the standard deviation, among other issues with the summary measures (Gelman et al., 2020).

Gelman et al. (2020) also offer the resolution of available-case analysis, in which different aspects of a problem are analysed using different subsets of a dataset. The issue is that the different analyses are based on different subsets of the data and may not be consistent with each other (Gelman et al., 2020). Additionally, this biases the analysis by ignoring the subset with *"missing"* values. A second resolution offered is nonresponse weighting, in which the inverse of predicted probabilities of response to the predicted nonresponse in a variable is used as a survey weight (Gelman et al., 2020). An issue with this method is that it becomes more complicated as there is *"missing"* data increases and standard errors can become erratic (Gelman et al., 2020).

Associated with the use of the mean to replace a *"missing"* value are other techniques of imputation, such as last value carried forward, using information from related observations, and imputation based on logical rules (Gelman et al., 2020). Last value carried forward means replacing the *"missing"* outcome variable with the last recorded pre-treatment value. While apparently conservative, this approach can have an anti-conservative bias as Gelman et al. point out with their HIV example. Using information from related observations means to replace the *"missing"* value with a value from an associated or similar variable (Gelman et al., 2020). However, this brings in the assumption that the variable from which the replacing value comes is representative of the variable in which the replaced value occurs. Which can be a difficult assumption to defend. Lastly, imputation based on logical rules means imputing *"missing"* values using some set of logical rules (i.e. people who refused to answer earnings and did not work and reported zero working months the previous year most likely have zero earnings) (Gelman et al., 2020). However, this may be difficult to apply in situations in which the *"missingness"* mechanism cannot be identified.

The last two methods of resolving *"missing"* to be discussed are multiple imputation and Bayesian imputation, the first as discussed in Gelman et al. (2020) and the second as discussed in McElreath (2019a). This is not to say that these are the only other two solution possible, just that they were the common ones mentioned in the literature beyond specific situations of *"missing"* data.

Multiple imputation is an imputation technique that create multiple imputations for each *"missing"* value, each of which is predicted from a different model and each of which reflects the sampling variability (Gelman et al., 2020). Effectively, multiple versions of the imputation value are generated from varying models and then the resulting imputations can be combined and then analysed. This technique has frequentist base and is commonly used in survey nonresponse (McElreath, 2019a). Gelman et al. (2020) and Wu and Thompson (2020) both present excellent examples on how to work through multiple imputation.

In his book, and online lecture, McElreath (2019a; 2019b) presents a resolution to *"missing"* values that involves a Bayesian approach. As McElreath says, "the statistical trick with Bayesian imputation is to model the variable that has missing values" (2019a, p. 516). Effectively the distribution of the observed values becomes the prior for the missing values, which when used with a model will produce a posterior for the *"missing"* values (McElreath, 2019a).

1.5 Considerations for imputation

While it may be necessary to impute values to be able to perform an analysis, considerations need to be taken. The act itself is replacing a value, the value being replaced is not *"missing"* it just does not look as expected and prevents an analysis. By replacing the value it is adjusting the data and the distribution. Care needs to be taken when doing so. Particular attention need be paid to the standard error. Lastly, when replacing survey values care needs to be taken that the imputed value be represntative of the respondent. A person should come first, not the analysis. It may be that the values to be imputed are not representative in that the value to be imputed cannot be calculated from or does not exist in the data in the first place.

2 Additional reading

Little, R. J. A., & Rubin, D. B. (2020). *Statistical analysis with missing data* (3rd ed.). Wiley.

While reading (and watching) the pieces in the references section (as well as those pieces in the preview review I wrote) I kept coming across the name Little and Rubin. Particularly, the book *Statistical Analysis with Missing Data*. I have been able to locate a copy of the book; however, I have not had a chance to read it yet. I wanted to add it here as an additional item that I believe would be useful to this work.

3 References

3.1 Gelman, A., Hill, J., & Vehtari, A. (2020). Chapter 17: Poststratification and missing-data imputation. *Regression and other stories* (pp. 313-333). Cambridge University Press.

3.2 King, G. [Gary King]. (2020, Aug 16). *18. Missing data* [Video]. YouTube. https://www.youtube.com/watch?v=qlPs8Ioa56Y&feature=youtu.be

3.3 McElreath, R. (2019a). Chapter 15: Missing data and other opportunities. *Statistical Rethinking* (2nd ed., pp. 499-532). CRC Press.

3.4 McElreath, R. [Richard McElreath].(2019b, March 1). *Statistical rethinking winter 2019 lecture 20* [Video]. YouTube. https://www.youtube.com/watch?v=UgLF0aLk85s

3.5 Wu, C., & Thompson, M. E. (2020). Chapter 9: Methods for handling missing data. In J. Chen & D. Chen (Eds.), *Sampling theory and practice* (pp. 193-219). Springer.