SAS GLOBAL FORUM 2020 #SASGF

Paper 5072-2020

Dealing with Missing Data in Epidemiological and Clinical Research

Andrew T. Kuligowski, Independent Consultant
Lida Gharibvand, Loma Linda University

ABSTRACT

Missing values can have a surprising impact on the way data is analyzed and processed. Epidemiological and clinical research typically involve complex data and large databases that frequently contain missing data. The impact of missing data on data analysis and research findings can be significant, so it is important to develop a sound methodology to deal with it. Fortunately, there are powerful tools to represent and reference the missing data in SAS® analytics. There are a number of SAS functions and procedures that enable differentiated approaches for handling missing data. However, dealing with missing data can still be a bit of a minefield. This paper presents an introduction to categories of missing data and demonstrates some techniques that researchers can use to deal with missing data.

KEYWORDS: Missing values, MCAR, MAR, MNAR, Complete case analysis Listwise deletion, Pairwise deletion, Mean Imputation, Regression Imputation, Stochastic Imputation

INTRODUCTION

Survey data frequently contain missing observations due to respondent refusal, errors in fieldwork, etc. Missing data are unavoidable in survey methods, epidemiological and clinical research, and business data. Large amounts of missing data can bias survey estimates and results. Many statistical techniques assume or require complete data, so missing data can reduce effective sample size and statistical power.

It is important to understand how SAS handles missing values when you execute statements. Depending on the statements being used, SAS might handle missing values in different ways. For example, it might treat a missing value as the lowest possible value (e.g., frequency tables in PROC FREQ), or it might omit the value from the computation (e.g., regression).

Internally, SAS treats numeric missing values as an extremely small number. Most of the time, the user will probably not be affected by this internal machination. In general, if you are subsetting data or doing any kind of conditional logic based on continuous numeric values, you should always explicitly tell SAS how to handle missing values.

METHODS:

DEFINITION OF MISSING VALUES

A *missing value* is a value that indicates that no data value is stored for the variable in the current observation. There are three kinds of missing values:

- numeric
- character
- Special numeric.

HOW TO REPRESENT MISSING VALUES IN RAW DATA

SAS represents missing data in multiple ways, depending on the type of value that is missing. The basic rule is that character values are represented by a blank (' ') or a null ('') and numeric values are

represented by a single period (.) “Representing Missing Values” (Table 1) shows how each type is depicted in SAS.

Missing Values	Representation in Data	
Numeric	.	a single decimal point
Character	' '	a blank (enclosed in single or double quotes)
Special	.letter	a decimal point followed by a letter, for example, .B
Special	._	a decimal point followed by an underscore

Table 1 - REPRESENTING MISSING VALUES

SAS allows for 27 *special missing values*, represented as a period followed by one of the letters A-Z or the underscore “_”. When sorting, the alphabetic versions of Special Missing Values are considered “larger” (in alphabetic order) than an ordinary missing value – but the Underscore is considered “smaller” than an ordinary missing value.

PROBLEMS

Researchers usually address missing data by only including complete cases in the analysis — those individuals who have no missing data in any of the variables required for that analysis. However, results of such analyses can be biased. Furthermore, the cumulative effect of missing data in several variables often leads to exclusion of a substantial proportion of the original sample, which in turn causes a substantial loss of precision and power.

CLASSIFICATION OF MISSING VARIABLE

- **Missing completely at random (MCAR)**

If the probability of being missing is the same for all cases, then the data are said to be missing completely at random (MCAR). This effectively implies that causes of the missing data are unrelated to the data. We may consequently ignore many of the complexities that arise because data are missing, apart from the obvious loss of information. An example of MCAR is a weighing scale that ran out of batteries. Some of the data will be missing simply because of bad luck. Another example is when we take a random sample of a population, where each member has the same chance of being included in the sample. The (unobserved) data of members in the population that were not included in the sample are MCAR. While convenient, MCAR is often unrealistic for the data at hand.

- **Missing at random (MAR)**

If the probability of being missed is the same only within groups defined by the observed data, then the data are missing at random (MAR). MAR is a much broader class than MCAR. For example, when placed on a soft surface, a weighing scale may produce more missing values than when placed on a hard surface. Such data are thus not MCAR. If, however, we know surface type and we assume MCAR with the type of surface, then the data are MAR. Another example of MAR is when we take a sample from a population, where the probability to be included depends on some known property. MAR is more general and more realistic than MCAR. Modern missing data methods generally start from the MAR assumption.

- **Missing not at random (MNAR)**

Finally, data are said to be missing not at random if the value of the unobserved variable itself predicts missingness. A classic example of this is income. Individuals with very high incomes are more likely to decline to answer questions about their income than individuals with more moderate incomes.

If neither MCAR nor MAR holds, then we speak of missing not at random (MNAR). In the literature one can also find the term NMAR (not missing at random) for the same concept. MNAR means that the probability of being missing varies for reasons that are unknown to us. For example, the weighing scale mechanism may wear out over time, producing more missing data as time progresses, but we may fail to note this. If

the heavier objects are measured later in time, then we obtain a distribution of the measurements that will be distorted. MNAR includes the possibility that the scale produces more missing values for the heavier objects (as above), a situation that might be difficult to recognize and handle. An example of MNAR in public opinion research occurs if those with weaker opinions respond less often. MNAR is the most complex case. Strategies to handle MNAR are to find more data about the causes for the missingness, or to perform what-if analyses to see how sensitive the results are under various scenarios.

Patterns of data loss are typically described as either ignorable or non-ignorable.

Types of ignorable missing data:

- Missing completely at random (MCAR): the missing observations on a given variable differ from the observed scores on that variable only by chance, and the missing observations are further not related to any other variable.
- Missing at random (MAR): the missing observations on a given variable differ from the observed scores on that variable only by chance.

Non-ignorable missing data:

- Missing not at random (MNAR): cases with missing data differ from cases with complete data for some reason, rather than randomly.

An understanding of the missing data mechanism(s) present in your data is important because different types of missing data require different treatments. When data are missing completely at random, analyzing only the complete cases will not result in biased parameter estimates (e.g., regression coefficients). However, the sample size for an analysis can be substantially reduced, leading to larger standard errors. In contrast, analyzing only complete cases for data that are either missing at random, or missing not at random can lead to biased parameter estimates. Multiple imputation and other modern methods such as direct maximum likelihood generally assume that the data are at least MAR, meaning that this procedure can also be used on data that are missing completely at random. Statistical models have also been developed for modeling the MNAR processes; however, these models are beyond the scope of this paper.

SOLUTIONS

COMMON TECHNIQUES FOR DEALING WITH MISSING DATA

In this section, we are going to discuss some common techniques for dealing with missing data and briefly discuss their limitations.

1. List-wise deletion

Complete case analysis (likewise deletion) is the default way of handling incomplete data in many statistical packages, including SAS, SPSS and Stata. The major advantage of complete cases analysis is convenience. If the data are MCAR, list-wise deletion produces unbiased estimates of means, variances and regression weights. It produces standard errors and significance levels that are correct for the reduced subset of data, but that are often larger relative to all data.
A disadvantage of list-wise deletion is that it is potentially wasteful.
The implications of missing data are different depending on where they occur and the parameter and model form of the complete data analysis. In context of regression analysis, list-wise deletion possesses some unique properties that make it attractive in a particular setting.

2. Pairwise deletion

This method involves estimating means, variances and covariances based on all available non-missing cases. Meaning that a covariance (or correlation) matrix is computed where each element is based on the full set of cases with non-missing values for each pair of variables. This method became popular because

the loss of power due to missing information is not as substantial as with complete case analysis. Below we look at the pairwise correlations between the outcome read and each of the predictors, write, prog, female, and math. Depending on the pairwise comparison examined, the sample size will change based on the amount of missing present in one or both variables.

As with Complete Case Analysis, this method will introduce bias into the parameter estimates unless the mechanism of missing data is MCAR. Therefore, this method is also not recommended.

3. Mean Imputation

This method involves replacing the missing values for an individual variable with its overall estimated mean from the available cases. While this is a simple and easily implemented method for dealing with missing values it has some unfortunate consequences. The most important problem with mean imputation, also called mean substitution, is that it will result in an artificial reduction in variability due to the fact you are imputing values at the center of the variable's distribution. This also has the unintended consequence of changing the magnitude of correlations between the imputed variable and other variables. We can demonstrate this phenomenon in our data.

4. Regression Imputation

Regression imputation incorporates knowledge of other variables with the idea of producing smarter imputations. A slightly more sophisticated type of imputation is a regression/conditional mean imputation, which replaces missing values with predicted scores from a regression equation. The strength of this approach is that it uses complete information to impute values. The drawback here is that all your predicted values will fall directly on the regression line once again, decreasing variability – just not as much as with unconditional mean imputation. Moreover, statistical models cannot distinguish between observed and imputed values and therefore do not incorporate into the model the error or uncertainty associated with that imputed value. Additionally, you will see that this method will also inflate the associations between variables, because it imputes values that are perfectly correlated with one another. Unfortunately, even under the assumption of MCAR, regression imputation will upwardly bias correlations and R-squared statistics.

5. Stochastic regression Imputation

In recognition of the problems with regression imputation and the reduced variability associated with this approach, researchers developed a technique to incorporate or "add back" lost variability. A residual term, that is randomly drawn from a normal distribution with mean zero and variance equal to the residual variance from the regression model, is added to the predicted scores from the regression imputation thus restoring some of the lost variability. This method is superior to the previous methods as it will produce unbiased coefficient estimates under MAR. However, the standard errors produced during regression estimation were less biased then the single imputation approach, will still be attenuated.

A problem with imputing only a single value for every missing value is that this does not reflect our uncertainty about the predictions. Standard errors may therefore be biased (too small). An alternative is to replace each missing value with multiple plausible values. This represents the uncertainty about the right value to impute. Data analyses from multiple imputed datasets can be combined to produce estimates and confidence intervals that incorporate missing-data uncertainty.

SUMMARY

Table 2 provides a summary of the methods discussed in this paper. The table addresses two topics: whether the method yields the correct results on average (unbiasedness), and weather it produces the correct standard error.

	Mean	Unbiased Reg weight	Correlation	Standard Error
Likewise deletion	MCAR	MCAR	MCAR	Too large
Pairwise deletion	MACR	MACR	MACR	Complicated
Mean imputation	MACR	-----	-----	Too small
Regression imputation	MAR	MAR	------	Too small
Stochastic imputation	MAR	MAR	MAR	Too small

Table2. OVERVIEW OF ASSUMPTIONS MADE BY SIMPLE METHODS

REFERENCES AND RELATED READING

Enders, Craig (2010). *Applied Missing Data Analysis*. New York, NY: The Guilford Press.

Johnson and Young (2011). Towards Best Practices in Analyszing Datasets with Missing Data: Comparisons and Recomendations. *Journal of Marriage and Family,* 73(5): 926-45.

Kuligowski, Andrew T. (2009) "The Ins and Outs and Ups and Downs of Sorted Data". *Proceedings of the SAS® Global Forum 2009 Conference*. Cary, NC: SAS Institute, Inc. http://support.sas.com/resources/papers/proceedings09/142-2009.pdf

SAS Institute Inc. (2017). *Base SAS® 9.4 Procedures Guide, Seventh Edition*. Cary, NC: SAS Institute Inc. https://documentation.sas.com/api/docsets/proc/9.4/content/proc.pdf

SAS Institute Inc. 2016. SAS® 9.4 Companion for Windows, Fifth Edition. Cary, NC: SAS Institute Inc.http://support.sas.com/documentation/cdl/en/hostwin/69955/PDF/default/hostwin.pdf

SAS Institute Inc. 2016. SAS® 9.4 Functions and CALL Routines: Reference, Fifth Edition. Cary, NC: SAS Institute Inc. https://documentation.sas.com/api/docsets/lefunctionsref/9.4/content/lefunctionsref.pdf

van Buuren , Stef (2018). Flexible Imputation of Missing Data. Boca Raton, FL:CRC/Chapman & Hall

This list is meant to whet the appetite. The reader is heartily encouraged to search for other fine papers on www.lexjansen.com

ACKNOWLEDGMENTS

The authors wish to acknowledge and thank SAS Global Forum 2020 Chair Lisa Mendez and her Content Advisory Team (CAT) for believing that the topic would be of interest to conference attendees, and that the submitters would be the right people to compose and present it to them.

CONTACT INFORMATION

In the event of any questions, comments, or whatever, you can contact the authors via email:

Andrew T. Kuligowski
KuligowskiConference@gmail.com

Lida Gharibvand
lgharibvand@llu.edu