Paper 319-2012

Knowledge (of your missing data) is power: handling missing values in your SAS® dataset

Theresa Schwartz, Rachel Zeig-Owens

Abstract

Before conducting any statistical tests it is important to check for missing values and evaluate how they may influence your study conclusions. This paper presents an overview of considerations that need to be made when confronted with missing data. We describe how to efficiently check for missing values, as well as investigate how SAS handles missing values and what can be done to correct for these missing values in your analysis.

Background

It probably does not come as a surprise to hear that SAS runs analytic procedures on only complete records. In order to be considered complete, a record needs to be non-missing for each dependent and independent variable included in the statistical model. This is what statisticians call complete case analysis. Complete case analysis makes things easier from a computational perspective, because it removes incomplete records, but eliminating missing data can lead to bias, loss of statistical power, and loss of valuable data collected. Depending on the project, a lot of time, money, and effort can be put into data collection. One of the last things a principal investigator wants to hear is that parts of the final data are not usable (and understandably so). Missing data can compromise the integrity of the study by affecting the three broad categories associated with the scientific method: reliability, validity and generalizability. The validity of conclusions drawn from incomplete data may be questionable as well as the reliability of the original study design. If the proportion of incomplete cases is high, it can limit the generalizability of the analyses because they represent only a subset of the study population.[1]

Luckily there are steps that can be taken to prevent missing data as early as the study design phase, in addition to format conventions that can be utilized during data entry to properly keep track of the type of missing values. As for unavoidable missing values, in order to prevent conflicts and biases, and to maximize data collected, it has become common practice to adjust for missing values in a number of ways. Widely accepted methods are multiple imputation (MI) and maximum likelihood. Before adjusting for missing values one must become familiar key assumptions of statistical analyses and determine the cost/benefit of adjusting for missing data points. You must also consider the reason values are missing, the fraction of cases missing and the extent the missing and non-missing values differ through patterns and other diagnostic measures. Assessing a missing data problem can often be considered more art than science. One must investigate the origin of missing values and understand the impact they will have on study conclusions.

We provide here a summary of steps that can be taken to prevent missing data problems in data collection/entry as well as evaluate missing data issues to check that analytical assumptions are met with the help of the SAS Missing Data Macro. Our discussion will focus on data obtained through self-administered surveys.

Data Collection

Proper coding for missing data and awareness of why it is missing is crucial at all stages of data collection and analysis. Survey data can be missing for a variety of reasons, some of which are beyond the researcher's control while others may be preemptively avoided with study design. One such issue is survey length. If a survey takes a long time to complete, a respondent may purposely skip questions or not complete the survey due to boredom, frustration, or fatigue. A way to avoid this, which has become common practice for electronic and paper based surveys, is to design question skip patterns. This can allow a respondent to skip entire sections based on his or her responses to a few key questions. With this survey design, missing data is planned in advance, assuming that the skip patterns are administered and programmed correctly.

A survey participant may also refuse or forget to answer an individual or set of questions. It may be impossible to avoid this altogether, but attention to the sensitivity of questions being asked may cut back on refusals. Designing question response options and therefore variable types that are categorical ranges or intervals instead of exact

values, when the exact amount is not necessary, can alleviate this problem.[1] For example, when collecting potentially sensitive information such as age, using intervals such as 18-25, 26-35, 40-45, and 46-55 instead of asking to provide exact age may lead to more responses. Similar caution can be taken with other demographic information such as income and race. Recognize that survey creation is a skill in itself, and as the study statistician or programmer your knowledge of how the data will be used in analyses is critical to survey design. Proper focus must be taken when designing not only variable type but also question wording and instructions.

Another common issue in data collection is missing data due to longitudinal loss to follow-up. When measurements are taken too frequently and within short time intervals this can lead to responder fatigue and frustration. Careful thought should be put into the number and time between follow-up measures and should be determined by not only the study question but the outcome measures being collected. There needs to be as much or as little time between surveys to see a possible change, without causing a burden on respondents. Often data collection time intervals are determined by other reasons such as convenience and this can lead to problems with retention.[1]

Lastly, data may be missing due to failure in equipment or methods during the data collection process or data entry. Establishing quality assurance practices throughout the data collection process is integral to catching problems as they arise instead of finding them after data collection has been closed!

Missing Values and SAS

Analytic Procedures

No matter how careful you are, during data collection, missing data may not be avoided altogether so it is important to understand how SAS handles any incomplete records. SAS performs analyses on only complete cases and uses two ways to eliminate missing data: listwise and pairwise. Listwise deletion removes any row or record that has at least one missing value. For example if you are running a prediction model and are missing any predictor for a participant (e.g. race) the entire row, representing all of that participant's data, including all predictors and outcome measures will be eliminated from the analysis. This is the default for how most SAS procedures handle missing values, including: PROC FREQ, PROC REG, PROC FACTOR, and PROC GLM. The second method, pairwise deletion, uses all available data and eliminates only pairs of variables that have missing values. By default PROC CORR produces correlations based on pairwise deletion and gives correlations for all valid pairs which are not missing any values. PROC FREQ and PROC CORR can be specified to use an alternate deletion method by adding a MISSING option. Please see SAS documentation for more details. Performing analyses on complete cases will lead to biased estimates if assumptions are not withheld. In cases where bias is inevitable and the default treatment of missing data cannot be altered in SAS analytical procedures, missing data adjustments should be considered.

Keeping in mind how SAS handles missing data, the next step is to be aware of how this will affect the statistical test you choose. For example, when comparing two models with different predictors to determine the best fit the usual statistical inference methods (e.g. maximum likelihood) for comparing models are only valid when applied to the same base data set. When a predictor in one model contains missing values, causing rows to be removed listwise, this data set is then altered and non-comparable to the data set used in the previous model that did not have any missing values in its predictors. In this case listwise deletion created a violation of a key assumption.

Special Missing Values

Because SAS treats all missing data equally deleting it either listwise or pairwise from procedures, it still may be important to your study design to keep track of what types of missing data you have. To designate types of missing values, some programmers and investigators have come into the habit of assigning known missing values as negative or really high and improbable values (i.e. 9999) so that in the analysis stage they can be filtered out. This can lead to computational error and incorrect conclusions if the data analyst is not properly informed of the coding scheme. When data is collected from survey responses and the respondent refused to answer, did not know the information or it was unavailable, or the respondent could not be reached to participate it may be necessary to differentiate these responses as they are informative for study and questionnaire design in the future. If we coded special missing values as a standard "." or blank, we would not be able to distinguish them from absolute or truly missing values. Special missing codes which are "." followed by a single capital letter and can be applied using a format. A common special missing code is ".R" for a subject's refusal to answer a question. In the PROC FREQ and other descriptive procedures these special missing values can be tabulated separately, however, they are all treated the same as an absolute missing in statistical and modeling procedures. Below is an example of a survey variable that has special missing codes in which respondents could chose "Don't Know" .D and "Refused" .R. Please see SAS documentation for more information.

When we run the PROC FREQ procedure on a variable with special missing codes (.D and .R) the responses are treated as absolute missing "." as in other SAS analytical procedures. This is shown as ODS output displayed in Table 1 and is called listwise deletion.

```
PROC FREQ Data=test1;
table race;
run;
```

Table 1. Special Missing Values: PROC FREQ Without MISSING Option

race	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	9067	96.25	9067	96.25
2	285	3.03	9352	99.28
3	68	0.72	9420	100.00

514 missing values

When we specify the MISSING option in PROC FREQ the special missing codes are separated out so we can view the exact number of each type of missing value.

```
PROC FREQ Data=test1;
table race/MISSING;
run;
```

Table 2. Special Missing Values: PROC FREQ With MISSING Option

race	Frequency	Percent	Cumulative Frequency	Cumulative Percent
D	307	3.09	307	3.09
R	207	2.08	514	5.17
1	9067	91.27	9581	96.45
2	285	2.87	9866	99.32
3	68	0.68	9934	100.00

Missing Data Mechanisms

There are several mechanisms that lead to missing data and consequently different missing data assumptions. Rubin was the first to describe missing at random (MAR) which states that the "missingness" of a single variable may depend on other observed variables in the model of interest but not on the variable itself after controlling for these other variables.[2,3] In other words, the mechanism behind a variable (X) being missing cannot be dependent on the value of X. An example of MAR for an income variable is those in a lower age group are more likely to answer the income question, regardless of their level of income. In this scenario, age predicts missing the income variable, but responses to the income variable itself do not predict it being missing. MAR is the primary assumption of MI methods.

A special and more restrictive case of MAR is missing completely at random (MCAR) which states that the "missingness" of a variable does not depend on any other variables or on the other possibly missing values of the variable of interest. In other words, the fact that the data is missing has nothing to do with the value the variable would have obtained or any other relevant variables in a model. MCAR might occur if respondents were randomly assigned to take a subset of questions regardless of answers to any other survey questions. MCAR is the primary assumption for complete case analysis and when withheld leads to unbiased estimates. However, although not impossible, MCAR situations rarely occur.[3]

It should be noted that a requirement of all missing data techniques that goes hand in hand with the assumption of MCAR and MAR is that missing data should be "ignorable." This does not mean ignorable in the sense that we disregard it altogether in analyses. Missing data is considered ignorable if the reason for it being missing is related to information that is known (as in MCAR and MAR), but does not directly affect the parameters being estimated in your model.[3] In other words, if the missing data is from a necessary and relevant predictor, it cannot be ignored. Also, if respondents fail to respond to survey questions because they refused to answer, this is not ignorable. However, if respondents fail to respond because they forget to attend a follow-up visit, this forgetfulness is unrelated to the construct being measured at the visit, and missing data can be considered ignorable.

Although it is not straightforward to test that the MAR or MCAR assumptions are being withheld, looking at the data can give some clues. When MAR and MCAR are violated data is said to be not missing at random (NMAR). An example of NMAR is if the probability of answering the income question from the example above was dependent on income itself, such as those with lower incomes were less likely to respond to the question. In the case of NMAR, the missing data mechanism is non-ignorable and missing data methods should be performed with caution. Data imputation methods such as MI and maximum likelihood parameter estimation are highly dependent on the MAR assumption. Therefore, if this assumption is violated these popular analytical methods should not be used. It is not possible, even in theory, to statistically adjust for a NMAR situation without additional data collection.[4] Methods that have been developed to analyze NMAR data are beyond the scope of this paper.

Missing Data Patterns

Unfortunately, there is no standard statistical test or method to determine if missing data is MAR. Despite no hard and fast rules, looking at the proportion of missing values, as well as analyzing the patterns of missing variables across survey subjects, can be very telling and help with diagnosing missing data issues. Knowing your data is necessary because you need to determine if the missing you observe is of concern or if it is expected based on your project and survey design. Having an idea of what variables you wish to analyze is important before investigating possible missing data issues. Looking at the entire data set for missing data patterns can be a waste of time in a situation where an investigator collects more information than he or she intends to analyze for a given hypothesis. Reviewing extraneous information could make the missing data problem seem more or less alarming than it really is. Once variables have been selected and you have begun to embark on missing data diagnostics, there are a few lines of questioning that could help determine the characteristics of the missing data. Not all pertain to all kinds of data sets or survey designs, but outlined below are a few questions that should always be considered:

First, look at missing variable proportions and ask: *What proportion of the data is missing for a particular variable?* If a large percentage is missing, follow-up questions could be: *Was it a sensitive information variable? Was it involved in skip logic?*

Second, when looking at missing data patterns one should also ask: *How many variables are missing for respondents? Are there a group of respondents missing virtually all of the variables? Could the non-missing variables be pre-existing demographic information?* (note: This will be discussed later as unit non-response).

Lastly, one should ask*: Are any variables missing in tandem? What is the relationship between these two variables? Do response variables differ for those missing and non-missing key variables?*

Running basic descriptive procedure such as PROC FREQ, PROC MEANS and PROC UNIVARIATE is a great start to looking at your data diagnostically at the beginning of the analysis process. By using the MISSING option in PROC MEANS all missing and special missing values in a class variable will be printed. To get a better picture of a missing data problem the proportion of missing values and missing data patterns should be evaluated. PROC MI will produce missing data patterns for variables specified, however, does not create missing variable indicators or produce proportions. In this paper we used the Missing Data SAS Macro created by researchers at Columbia University to produce Tables 3-7.[5] The macro is very user friendly and was created to easily look at missing indicator patterns, missing data proportions, concordance of missing data and unit non-response by creating diagnostic work data sets. It should be noted that by default the macro does not treat special missing codes as absolute missing, however, this is easily fixed by altering one line of the macro to include the specific special missing codes in your data set. With the alteration the macro will read in all specified special missing values as absolute missing values just as any SAS procedure does. Further details on the macro can be found in the SAS Global Forum paper, however, we have supplied the specific macro calls with the tables to give a better idea of its utility.

The proportion of subjects missing key variables should be evaluated as seen by Table 3 which was created by the missPattern2 of the Missing Data SAS Macro. Variables with a high proportion of missing values should be checked for any possible reason behind the missing. There is not an agreed upon cutoff for the proportion of a variable that

should be missing to be considered alarming. This is data dependent, as 25% missing may be alarming for a demographic variable but not so for a question involved in skip pattern logic. The proportion should also be considered when looking into methods such as MI to make sure there is enough complete information to make valid inferences about missing variables.

%macro missingPattern(**datain=work.example**, **varlist=occupation race exposed income age ever_smoke**, exclude=, missPattern1=, dataout1=, **missPattern2='TRUE', dataout2= work.Table3**, missPattern3=, dataout3=, missPattern4=, dataout4=);

Table 3. Number and Proportion of Variables Missing

var	num_miss	prop_miss
occupation	0	0
race	0	0
exposed	0	0
income	323	3.2782
age	1367	13.8739
ever_smoke	1386	14.0668

Missing data patterns are a great way to see how missing and complete values relate to each other and in turn can give clues as to whether the MAR or MCAR assumptions have been violated. These patterns are visualized by creating dummy variables that are coded as "1" when the variable is missing and "0" when it is observed. The indicator variables presented by subject can be useful to investigate patterns if there are patterns to be seen. Missing data can be missing for an individual variable in a univariate fashion, or across multiple variables. Patterns can be defined to be either structured (evident pattern) which have a mechanism behind missing values or unstructured (arbitrary with no pattern) which are more likely to be random and without a mechanism. To believe that your data are truly MCAR you do not want to see a structured pattern and would like to see each row of your data set to have either no missing variables or different, scattered missing data patterns. A few patterns to be aware of, and may appear clearly in your output are: monotone, unit nonresponse and file matching.

Monotone missing data (and close to monotone data patterns) can be evidence of longitudinal loss to follow-up. An example of monotone missing data is seen in Table 4, created by missPattern1 of the macro. This table shows missing indicators (1=missing) for select variables from a larger analysis data set. We see that the "current smoker" indicator variables (m_smoke_yr1 – m_smoke_yr3) show that data progresses to be missing more over time or in other words the variables in year 1 are more widely observed than in years 2 and 3, and data is progressively missing from one row to the next.

%macro missingPattern(**datain=work.example2, varlist=occupation race exposed smoke_yr1 smoke_yr2 smoke_y3**, exclude=, **missPattern1='TRUE', dataout1= work.Table4**, missPattern2=, dataout2=, missPattern3=, dataout3=, missPattern4=, dataout4=);

Table 4. Possible Evidence of Monotone Missing Data Pattern

m_occupation	m_exposed	m_RACE	m_smoke_yr1	m_smoke_yr2	m_smoke_yr3	N Obs	missPattern_prop
0	0	0	0	0	0	8354	80.5082
0	0	0	1	1	1	132	1.3397
0	0	0	0	1	1	113	3.1469
0	0	0	0	0	1	931	5.4489

Unit non-response occurs in survey data when an individual or group of individuals fail to complete any of the survey questions but may have preexisting demographic or identifying variables already in the data set. This is shown in row 2 of Table 5. In the data used to create Table 5 the survey consisted of 5 questions with 5 corresponding missing indicator variables (m_Q1 – m_Q5) linked automatically to an ID variable. Those who did not answer questions 1-5, but still had an ID variable non-missing were unit non-responders. Be careful not to assume non-response for patterns that have complete data to start then entirely missing variables as the items go forward. This could be evidence of participant fatigue or survey failure and not evidence of unit non-response. It is only non-response if the non-missing questions did not require participants to answer a question or fill in any information at time of the survey. In addition to looking at the pattern in Table 5, the macro can be called to look specifically for unit non-response with

the missPattern4. If there is possible unit non-response, the pattern will print as a data set as in Table 6, however, if there is no possible unit non-response, a message alerting to this will be printed in the log.

%macro missingPattern(**datain=work.example2**, **varlist=ID Q1 Q2 Q3 Q4 Q5,** exclude=, **missPattern1='TRUE', dataout1= work.Table5**, missPattern2=, dataout2=, missPattern3=, dataout3=, missPattern4=, dataout4=);

Table 5. Possible Evidence of Unit Non-response (SAS Missing Data Macro pattern 1)

m_ID	m_Q1	m_Q2	m_Q3	m_Q4	m_Q5	N Obs	missPattern_prop
0	0	0	0	0	0	8354	81.5082
0	1	1	1	1	1	132	1.3397
0	0	0	1	0	1	113	1.1469
0	0	0	0	0	1	931	9.4489
0	0	0	1	1	1	323	3.2782
0	0	1	0	1	1	323	3.2782

To look specifically for possible unit non-response:

%macro missingPattern(**datain=work.example2**, **varlist=ID Q1 Q2 Q3 Q4 Q5,** exclude=, missPattern1=, dataout1=, missPattern2=, dataout2=, missPattern3=, dataout3=, **missPattern4='TRUE', dataout4= work.Table6**);

Table 6. Possible Evidence of Unit Non-response (SAS Missing Data Macro pattern 4)

Obs	ID	m_Q1	m_Q2	m_Q3	m_Q4	m_Q5	N Obs	Num_var_missing
1	0	1	1	1	1	1	132	5

The third pattern is file matching. File matching occurs when variables are never observed together. It is important to be aware that when estimating the association between two variables that are never observed together that some of these parameters will be not estimable from the data. Knowledge of this is especially useful when looking at variables to include in MI models. If a variable is never observed with the variable to be imputed, it should not be used for imputation. File matching cannot always be detected by a pattern and may be more easily detected by looking at the occurrence of variables side by side. This is made a little easier by missPattern3 of the Missing Data SAS macro as shown in Table 7 below. This table looks at the full concordance of all of the variables specified by printing the following five percentages:

1. P00: % of subjects with both var1 and var2 observed;
2. P01; % of subjects with var1 observed but var2 missing;
3. P10: % of subjects with var1 missing but var2 observed; and
4. P11: % of subjects with both var1 and var2 missing.
5. The summary measure prop_concordance (= P00 + P11) presents the % of data that var1 and var2 are missing or observed together.

Note: This is only a subset of the actual table output by the Missing Data SAS Macro missPattern3, with file matching showing as a possible problem between variables occupation and exposed in line 2.

%macro missingPattern(**datain=work.example2, varlist= occupation race exposed income age ever_smoke,** exclude=, missPattern1=, dataout1=, missPattern2=, dataout2=, **missPattern3='TRUE', dataout3= Table7**, missPattern4=, dataout4=);

Table 7. Proportions of Select Variables Observed Together

var1	var2	P00	P01	P10	P11	prop_concordance
occupation	race	100	0	0	0	100
occupation	exposed	0	100	0	0	0
race	exposed	100	0	0	0	100
age	ever_smoke	84.786	1.3397	1.1469	12.7271	97.513
occupation	income	96.722	3.2782	0	0	96.722

Multiple Imputation (MI)

To correct for missing values and perform statistical analyses on a more "complete" data set it is common to use imputation. MI is a more superior method to single imputation because it takes into account the uncertainty of what the true values of the unknown data would be. In simple imputation a variable is imputed based on either the mean of the observed values or the mean conditional on the other variables included in your imputation model. Multiple imputation using PROC MI, creates multiple data sets by imputing a number of plausible values (the number being specified by the analyst) for the missing variables, taking into consideration the uncertainty of the true value of those missing. Statistical analyses are then run on these imputed data sets separately, using PROC MIANALYZE to combine the results into final parameter estimates. SAS documentation explains this procedure in further detail.

Using the diagnostic tools mentioned above is helpful when using PROC MI and PROC MIANALYZE. However, it should be noted that when MAR cannot be verified by the data that it becomes more credible with more variables being added to the imputation model.[6] Consideration should be made as to the observed proportion of variables being imputed, how many imputations should be made, what variables should be included in the imputation model, and which imputation method is most appropriate. An evident pattern or lack of pattern (which is an arbitrary missing data) will determine which imputation method is used. The PROC MI procedure allows for different methods: regression, propensity score, and Markov Chain Monte Carlo (MCMC). A parametric regression method that assumes multivariate normality or a nonparametric method that uses propensity scores is most suitable for a monotone missing data pattern. However, for an arbitrary missing data, the MCMC method that assumes multivariate normality would be appropriate. An extensive paper by Yang C. Yuan on multiple imputation and the available options and syntax can be found at SAS support. New to SAS 9.3 is an experimental statement that allows for an additional imputation method: fully conditional specification (FCS). FCS can be used with missing data that has an arbitrary missing data pattern with the advantage that it does not require as many iterations as MCMC.

Conclusion

Missing data is a complex issue that should be considered at every stage of study design, collection, and analysis. Properly keeping track of missing data is vital. Before analyzing any data you must come to an understanding as to what is missing, how SAS will treat these missing values, and how your statistical methods and study conclusions will be affected. Missing data diagnostics are necessary not just for knowledge of what data you have available to analyze, but to make an informed decision if statistical methods should be used to impute missing values, and if so, what method of imputation should be used.

References

1 McKnight PE, McKnight KM, Sidani S, et al. Missing Data, a gentle introduction. New York: Guilford Press, 2007

2 Little RJA, Rubin DB. Statistical Analysis with Missing Data. Second ed. New Jersey: John Wiley & Sons Inc., 2002

3 Allison PD. Missing Data. In: Millsap RE, Maydeu-Olivares A, eds. The SAGE Handbook of Quantitative Methods in Psychology
London: SAGE Publications, 2009

4 Wang C, Hall CB. Correction of bias from non-random missing longitudinal data using auxiliary information. Stat Med 2010; 29:671-679

5 Schwartz T, Chen Q, Duan N. Studying Missing Data Patterns Using a SAS® Macro. SAS Global Forum. Las Vegas, NV, 2011

6 Schafer JL. Analysis of Incomplete Multivariate Data: Chapman & Hall, 1997

Acknowledgments

The authors would like to acknowledge Beth Bruder, Cristin Casazza, Michelle Glaser, Charles Hall, and Jessica Weakley for their comments.

Contact Information

Theresa Schwartz
9 MetroTech Center, 5th floor
Brooklyn, NY 11201
email: Schwart@fdny.nyc.gov
phone: 718-999-5167