Winter 2014, Volume 4, Issue 4

THE STATSWHISPERER

The StatsWhisperer Newsletter is published by Dr. William Bannon and the staff at StatsWhisperer.™ For other resources in learning statistics visit us on the web at: www.statswhisperer.com

How to Do Perform Multiple Imputation (MI) Using SPSS

As the field of quantitative research evolves it is less and less acceptable to ignore missing data points in statistical analysis. In other words, suppose you have a sample of 1,000 study participants, of whom 800 provided complete data and 200 have some missing data. This situation would present you with the challenge of performing analysis among a sample with 200 study participants with missing data. In the past, many data analysts might solve this challenge of accounting for missing data by performing analysis using only the 800 study participants with complete data, while excluding (in essence ignoring) the 200 with some data missing. However, this practice is growingly regarded as unacceptable and analysts are often asked to employ methods to estimate missing data points in order to include the full sample in analysis.

At the same time, the there is greater recognition that the more traditional methods of accounting for missing data values, such as mean substitution, are generally flawed. That is to say, these methods often distort study results and are therefore unacceptable. This has led to a need among data analysts to become fluent in the latest sophisticated methods used to account for missing data points, such as *Multiple Imputation* (MI). However, while there is a great deal written on MI, there are not many resources for learning how to perform MI, especially in more common statistical software programs such as SPSS.

INSIDE THIS ISSUE

Subsequently, the current newsletter edition will present the nuts and bolts regarding how to perform MI using the SPSS statistical software program. Of course, the reader would also benefit from a more detailed discussion regarding the MI procedure itself and the remarkable manner in which MI estimates missing data values. However, the current newsletter will focus upon the one element that seems so hard to find, i.e., direction regarding how to actually perform the procedure. In other words, this issue is meant to be a key part of the puzzle for those individuals that are seeking to learn and employ MI toward accounting for missing data values within their dataset.

Performing Multiple Imputation (MI) Using SPSS

MI is regarded as an appropriate tool for estimating missing data values in the presence of study participant non-response or dropout (Bannon, 2013). Specifically, MI may be employed in a cross-sectional study design, as well as a longitudinal study where there is study participant dropout and the analyst would like to retain the sample size in an *intention to treat* style analysis. It is worth noting that many statisticians consider that a small percentage of study participants, generally around 5% in a somewhat large sample, may be removed from the analysis due to missing data values. Thus, if your sample is of a fair size and you have less than 5% of cases with missing data, you might consider excluding these cases from your sample as a means of addressing missing data values rather than using MI.

Briefly, MI runs simulations upon the data that are available as a means of estimating (i.e., replacing) the missing data value(s) with the values that would be the most likely response. Essentially, MI looks at patterns within the data and makes a probability judgment regarding what the most likely response would be. For example, if a study participant does not report his or her annual income, MI would consider how he or she responded to other items, as well as how other study participants of a similar profile responded to these items as well as their income level, and estimate what the missing annual income value probably would have been. Here are our steps:

Step 1: Assessing Patterns of Missingness

The first step in the MI process is to assess if there are any pattern to the missingness of the data, as data with and without patterns to the missingness are often treated in different ways.

When using MI in SPSS, we will employ a method that will identify if the data are missing in a random or systematic manner. SPSS will then use a different type of procedure in MI based upon these results. To begin this process go to:

Analyze, Multiple Imputation, Analyze Patterns

Within the dialogue box that opens, by default the first three boxes under the section labeled *Output* (i.e., Summary of Missing Values, Patterns of Missing Values, & Variables with the Highest Frequency of Missing Values) should have checks (if not please check them). Next:

In the left hand column, select (highlight) all the variables in the dataset, EXCEPT THE ID VARIABLE, and transfer them into the right hand column (labeled *Analyze Across Variables*) using the middle arrow in the dialogue box.

Within that dialogue box, note the line that reads *Maximum Number of Variables Displayed*. The default number is 25, so if you have more than 25 variables entered into the right hand column labeled *Analyze Across Variables*, please raise the number within that line (i.e., if you have 50 variables in the right hand column change the number on that line from 25 to 50). Beneath that line, please note the line that reads, *Minimum Percentage Missing for Variables* to be displayed, which tells SPSS to analyze variables with only a certain percentage of data missing, with the default value being 10%. However, since we want to examine all the missingness, change that value to 0.01%. Next click *OK (or Paste to save the syntax)*.

Performing Multiple Imputation (MI) Using SPSS (Continued)

We can now note the output produced from the Multiple Imputation analysis of missing data. At the top of the output you will see three pie charts that display the *Overall Summary of Missing Values* (this is the title over the charts).

The first pie chart on the left, labeled *Variables*, will display *how many/the percentage of variables that have missing data*.

The center pie chart, labeled *Cases*, will display *how many/the percentage of cases that have missing data* (i.e., cases that have at least one data value missing).

The final pie chart on the right, labeled *Values*, will display *how many/the percentage of values that are missing overall within the dataset*.

Next, below the pie charts, reference the table labeled *Variable Summary*. This chart lists the percentage of data missing within each study variable. The variables will be listed from highest to lowest regarding the percentage of missing data.

Now reference the next chart down, labeled *Missing Value Patterns*, which will be composed of small red and white rectangular lines. **This box will convey if there is a pattern to the missing data.** The X-axis lists the variables in the dataset left to right ordered from the variables with the least amount of missing data to the most.

We are examining this image for *monotonicity*, which is a rigid pattern of missing data within the figure, which will be displayed by a pattern within the rectangular red lines. For example, a pattern might be indicated by a concentration (bunching) of red lines in the upper left hand corner and a concentration of red lines in the lower right hand corner. However, no pattern to the missingness might be assumed if the red lines look evenly and randomly dispersed (not bunched) throughout the image. In other words, one could assume that there is not pattern to the missingness if there are islands of red among the white lines overall. We will make different decisions in our analysis based upon our conclusion that the data appear to be missing at random or not.

The final image labeled *Variable*, is the *Missing Value Pattern* frequency graph. Within the figure, the first bar on the left will likely be the largest indicating that the most prevalent pattern is the one *where no missing values are present*. This indicates that the most common pattern seen is that there are no data missing across all the variables.

The remaining patterns to the right will likely be considerably smaller. Also, if these patterns are also similar in size, the image will have indicated the patterns of missingness across the variables is rather consistent. In other words, there is no dominant pattern to the missingness that would be cause for concern.

In the majority of datasets, at least in my experience, the data tend to be missing at random. Subsequently, we will assume in our discussion here that this is what is indicated.

Step 2: Setting Parameters for MI

Our next step will be to give SPSS some parameters for conducting MI. To do this go to:

Transform, Random number Generators

Within the dialogue box, check the box that reads *Set Active Generator*. Then select the button for the *Mersenne Twister* (this is a

Performing Multiple Imputation (MI) Using SPSS (Continued)

random number generator program). Next, click *Set Starting Point*, then select the button for *Fixed Value*. You may keep the default value (20,000,000) for the *Fixed Value* as is. Then click *OK* (or *Paste* to save the syntax). Now we are ready to conduct the actual MI procedure.

Step 3: Conducting MI

To conduct the MI procedure, go to:

Analyze, Multiple Imputation, Impute Missing Data Values

In the dialogue box that appears, select (highlight) all the variables with missing data values in the left hand column and transfer them into the right hand column labeled *Variables in Model*. In other words, if all the variables have at least one datapoint missing, put them all over into the right hand column (except for the ID variable).

Next, note the line that reads *Imputation*. This line is used to instruct SPSS on the number of iterations to perform toward estimating the missing data values. The default number is 5, which means SPSS will estimate the missing values 5 times before producing a final estimate. Often for most purposes, we can leave the default number of 5 as a specification. In the fifth (final) estimate the values are averaged together in order to account for the variance in estimates. The procedure is known as *Multiple Imputation* because the final values are based upon multiple estimates, from which a single value is imputed.

Next, make sure the button for *Create a New Dataset* is selected. Then in the empty box under that line (labeled *Dataset Name*) *give a name* to the new dataset that will be produced that includes the MI estimates of the missing data values, we will use the name *Sample Imputed Dataset*.

Next, within that dialogue box, click the tab labeled *Method*. On this page you may want to select the *Automatic* method, as this method scans the data for monotonicity as we did earlier. If monotonicity is indicated the *Monotone* method will be employed, otherwise SPSS will default to the *Fully Conditional Specification* method (You can also select the *Custom* button if you would like to specify one of these methods yourself). In other words, you can let SPSS decide the best method to use.

You will see toward the bottom that reads *Model Type for Scale Variables*, where you can select the procedure used to estimate missing values. The default and most popular method used is linear regression.

Next, click the tab labeled *Constraints*. Then click the *Scan Data* button, which will give us the descriptives (e.g., % missing) for each of our variables in the box labeled *Variable Summary*. In this tab we will then **tell spss the valid range of responses for each variable, so the program does not estimate values that are outside the valid range of responses.** For example, if the range of responses for an item is 1-5, we do not want SPSS to have the option to estimate 6 (which is outside the range of 1-5) as an estimated value.

Next, scroll through the variables in the box labeled *Variable Summary*. As you scroll, note if the values for the *Observed Minimum* and *Observed Maximum* are outside the valid range of responses for any study variables.

If you find values outside the valid range, go to the next box labeled *Define Constraints*. Then locate the respective variable and enter the valid range of responses for that item.

Performing Multiple Imputation (MI) Using SPSS (Continued)

Keep in mind that you may only want to specify a minimum and not a maximum. For example, in the case of variables such as *annual income* or *number of times married*, zero may be a minimum score, but there may be no preset maximum. So here it would be appropriate to enter 0 as a minimum value and leave the maximum value blank.

You will see in the last column labeled *Rounding*, you can tell SPSS to round to a value to a certain number if desired. You will also see within this box a column labeled *Role*, where you can define the role of the variable. When using the linear regression method (as selected earlier) the default selection will be *Impute and Use as Predictor*, which you may choose to leave as is, as the specification should not have a great bearing on the procedure.

Next, select the tab labeled *Output*. Under the line titled *Display*, be sure to check the boxes for *Imputation Model* and *Descriptive statistics for Variables with Imputed Values*. Also, under the line *Iteration History*, click the box for *Create an Iteration History*. This will allow you to see the values for each of our 5 estimates one iteration at a time. You may find it is interesting to view how the first iteration may differ from the following iterations up to the fifth, which presents the final pooled value used for the missing data points.

There is another box here under the label *Create a New Dataset* where you must enter a dataset name. Here you may want to enter a phrase such as *Iteration History*, so you can easily identify where this history is located. Then click *OK (or Paste to save the syntax)*. SPSS usually takes a few minutes to conduct MI, but as you wait in the lower right hand corner of the screen you should see feedback reflecting that SPSS is running the various iterations for the MI process.

When SPSS is done running the MI process, you will have two new datasets. For example, you will have the dataset we names *Sample Imputed Dataset*, as well as the second one we named *Iteration History*.

In our new dataset named *Sample Imputed Dataset*, you will see a new variable in the far left hand column. This variable will identify the iterations contained within the dataset. Where the values are zeros (0), the data are the original values with the missing data. If you scroll down, this new variable will have ones (1), which indicates that this is the first iteration of imputed values. You can continue scrolling to see each iteration up to the final fifth (5) iteration. There will also be a small white and yellow button in the upper right hand corner that will allow you to go to each iteration more quickly (without scrolling).

Within each iteration (after the original data), you will see a series of yellow shaded cells, which are the cells that were originally empty (had missing data), but now have imputed (estimated) values.

When you approach doing inferential analysis using the imputed values, you will find that many tests are compatible with the MI procedure (when you go to the statistical procedure, compatible tests will have a cube with a swirl in it), while some are not available for analysis with these estimated values. When you conduct a statistical test, such as a t-test (which is usable with the MI values), within the output you will see several t-tests are presented. Specifically, SPSS will provide a t-test based upon the original data, all iterations, then a final t-test based upon the pooled estimates (where all data are present).

The Four Checks of Data Integrity (continued)

Of course, the point of conducting MI would be to use the results presented within the final pooled data version within the output. However, you may find it quite interesting to notice how the results of each t-test might differ among each iteration used in the t-test. For example, you might find some present statistically significant findings while others do not.

From here you can perform all other needed statistical analysis, as well as examine the rest of the output which describes what occurred during the imputation process, descriptive statistics, and the Iteration History database.

Final Comments

In this newsletter issue we reviewed the valuable process of conducting MI in SPSS. However, what we did not discuss is how using such a procedure might clash with one's ideological beliefs. For example, when we conduct a quantitative study, we mean to examine the data provided by study participants. However, when we use MI to estimate missing data points, to the degree that the data are imputed, we are not examining data provided by study participants, but data provided by SPSS.

To be sure, when we use a method of estimating missing data values, such as MI, we are gaining some things, such as a fuller sample and perhaps a more representative sample. However, we are actually giving up things too, such our confidence that we are analyzing the actual responses given by study participants.

Thus, there is a very strange catch 22 here. Specifically, we live in a world of quantitative analysis where ignoring missing data is understandably not acceptable. However, the methods used to account for missing data values only provides us with the best guess of what study participants would have probably responded. Consider how we noted that the t-test using only the data supplied by study participants might be statistically significant, but the same test with the MI values might not be. We are left only to wonder, what is the truth?

Some people are so curious about the effectiveness of MI that they will delete half the values within a dataset then use MI to estimate the deleted values. By comparing the original deleted values with the MI estimated values they can get an idea if how effective the MI procedure is.

In any case, whether we accept or do not accept these methods of account for missing data values, at least knowing how to perform MI may give us more of a choice.

REFERENCES

Bannon, W. M. (2013). *The 7 Steps of Data Analysis: A Manual for Conducting a Quantitative Analysis*. New York: StatsWhisperer Press. Only available at:

http://www.statswhisperer.com/books/

Thanks for your interest in our newsletter!

Contact us at: wb@statswhisperer.com with any questions or for information on:

Subscribing to this newsletter visit:

http://www.statswhisperer.com/newsletter/

For books visit:

http://www.statswhisperer.com/books/

For webinars visit:

http://www.statswhisperer.com/webinars/