Paper 271-28

An Alternative to PROC MI for Large Samples

David Lanning, State Farm Insurance Companies, Bloomington, IL
Doug Berry, State Farm Insurance Companies, Bloomington, IL

ABSTRACT

Often, when building statistical models, you need to modify or recode data that have missing values. There are several accepted methods to do just that, including the new SAS/STAT procedure PROC MI (Multiple Imputation). PROC MI creates multiple, if desired, imputed data sets for incomplete multivariate data. After the data sets are created, another procedure, PROC MIANALYZE can be used to generate valid statistical inferences by combining the results from the multiple imputed data sets. PROC MI works well for samples of a few thousand records or less with a limited number of variables. Sometimes, though, researchers have tens of thousands, or even millions of records and a large number of variables. In these cases, PROC MI can become slow, or not even be able to impute the data at all. This paper explains how to accomplish an alternative to PROC MI for large datasets using PROC MEANS, SAS Macros, Arrays, and IF/THEN logic in the Data Step.

INTRODUCTION

As statistical analysts, we often wonder if the methods we use to manipulate our data are correct. Will using mean substitution or median replacement skew my results? Will capping an age at 85 due to an error in data input cause more problems? Listed below are examples of several methods used to deal with missing data.

Many SAS statistical procedures exclude observations with any missing variable values. This provides simplicity by only using complete cases. However, this may, in fact ignore important differences between the complete and incomplete cases, creating inferences not applicable to the entire population. Another concern with this method is the possibility that a missing value exists in every observation, thus giving you nothing to work with.

Other procedures, such as the PROC CORR use all the cases with available information by estimating a variable mean for each variable based on the nonmissing values of the variable. PROC CORR also estimates a correlation matrix by using all nonmissing values for a pair of variables. This method uses all existing data, but the resulting matrix may not be a positive definite.

Probably the most common method used to replace missing values is simple imputation. This method allows the data analyst to substitute nonmissing values where missing values used to exist. A simple example would be computing the median of a variable with all nonmissing values and then substituting the median for all missing values of that variable. This would allow the analyst to use all of the observations for that variable in the analysis. However, this method does not always reflect the uncertainty of the data.

The newest method of missing replacement is to use the PROC MI procedure to create multiple imputations for missing data. This method tries to represent a random sample of the missing values by creating multiple datasets for incomplete data. The result of this procedure lets the analyst use all of the observations in a dataset while maintaining the variability of the nonmissing data.

THE IDEA

PROC MI now offers a new alternative to the established ways analysts deal with missing data. PROC MI in conjunction with PROC MIANALYZE lessens the reality that accounting for the variability in the data biases the estimated variances or resulting inferences.

Data analysis today often involves large amounts of data and/or large numbers of variables. Unfortunately, while PROC MI is extremely capable in what it does regarding multiple imputations, it cannot handle the larger data sets that are so common in medium to large businesses today.

Because of this limitation within PROC MI, the idea was hatched to create a random replacement value for every missing value of all numeric variables in a given data set. The replacement values for a given variable are based on the nonmissing minimum and range of that variable.

The process involves using PROC MEANS to calculate the nonmissing minimum and range for every numeric variable in a data set. The minimum and range values are then feed, via a SAS Macro, into a calculation where a random uniform number is multiplied by the range and then added to the minimum. This calculation keeps the random number generated between the minimum and maximum values of each variable granted they were calculated prior to imputation. The result is the new imputed value that is inputted into the data using the IF/THEN statement in a data step. Simple statistics can then be used to show that the minimum, maximum and mean do not significantly change, thus aiding in further analysis.

CONCLUSION

There are many methods or processes an analyst can use to replace missing data. Each method holds merit, but PROC MI does the best job of randomly replacing missing data so the resulting statistics reflect the uncertainty due to the missing values. However, this procedure may take abnormal amounts of time to process results if the data being used is quite large or the numbers of variables are extensive. Hence, the alternative allows for faster processing while maintaining similar results that hold statistical value.

An added note, with any analysis, the researcher needs to determine how much missing is too much. Guidelines vary with respect to the research, but some variables may need to be excluded from the analysis if the variable has excessively large quantities of missing data. Another concern may be outliers. Obviously outliers can complicate an analyst's work and/or results. Knowing the data becomes a key component in determining each variables merit. This alternative is no different.

THE CODE

```
/****************************************/
/*****This data is for illustration*****/
/*************purposes only*************/
/****************************************/

data one;
  input abbrid $ cnt amt_rnd
        amt_dec age @@;
  datalines;
DTL 1 110 30.67 21      KRL 2 320 24.56 22
ARL 3 330 . 23          GLM . . . .
CRM 5 150 45.32 .       MGM 6 . 28.12 26
PRG . 270 . 27          RDS . . 34.00 .
JKS 9 . . .             DTL 2 160 54.98 20
KRL 4 180 67.54 22      ARL 6 140 34.13 24
GLM . . . .             CRM 8 170 23.87 .
MGM 2 . 43.76 26        PRG . 130 12.54 28
RDS . . . .             JKS 5 . . .
DTL 3 190 43.67 23      KRL 7 110 32.87 26
ARL 3 120 34.34 29      GLM . . . .
CRM 5 130 54.12 .       MGM 2 . 76.43 21
PRG . 140 19.92 25      RDS . . . .
JKS 8 . 53.13 .         JTS 6 . 73.93 30
;
run;

/****************************************/
/****randomly inputs valid data into****/
/******missing data values based on*****/
/*************range and min*************/
/****************************************/

proc means data=one noprint;
   var _numeric_;
   output out=three (drop=_freq_ _type_)
          min= range= / autoname;
run;


data _null_;
   set three;
   array ass{*} _numeric_;
   do i=1 to dim(ass);
    call sympput(vname(ass(i)),ass(i));
   end;
   drop i;
run;
```

```
data four;
   set one;
   array wit{*} _numeric_;
   do i=1 to dim(wit);
    if wit(i)=.
     then
      wit(i)=round((ranuni(0)*
      (symget(vname(wit(i))||'_range')))+
      (symget(vname(wit(i))||'_min')));
   end;
   drop i;
run;

proc means data=one min max range mean;
  title 'Original dataset with missing values';
run;
```

```
     Original dataset with missing values
             The MEANS Procedure
  Variable  Minimum   Maximum   Range    Mean
       cnt        1         9       8    4.58
   amt_rnd      110       270     160  163.00
   amt_dec    12.54     76.43   63.89   41.47
       age       20        30      10   24.56
```

```
proc means data=four min max range mean;
  title 'New dataset using random generated
         values for missing observations';
run;
```

```
     New dataset using random generated
       values for missing observations
             The MEANS Procedure
  Variable  Minimum   Maximum   Range    Mean
       cnt        1         9       8    4.57
   amt_rnd      110       270     160  173.86
   amt_dec    12.54     76.43   63.89   41.53
       age       20        30      10   25.46
```

References

Yuan, Y.C. (2002), "Multiple Imputation for Missing Data: Concepts and New Development," SUGI27 Conference Proceedings.

ACKNOWLEDGMENTS

I am grateful to Doug Berry for his valuable contributions to this creation and to his assistance in the preparation of this document.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

David Lanning
State Farm Insurance Companies
One State Farm Plaza – SC-3
Bloomington, IL 61710-001
Work Phone: (309)735-2723
Email: david.lanning.lyus@statefarm.com