Paper 1543-2014

Combining Type-III Analyses from Multiple Imputations

Binhuan Wang, New York University School of Medicine; Yixin Fang, New York University School of Medicine; Man Jin, Forest Research Institute

ABSTRACT

Missing data commonly occur in medical, psychiatry, and social researches. The SAS® procedures MI and MIANALYZE are often used to generate multiple imputations and then provide valid statistical inferences based on them. However, MIANALYZE is not applicable to combine type-III analyses obtained using multiple imputed datasets. In this manuscript, we write a macro to combine the type-III analyses generated from SAS procedure MIXED based on multiple imputations. The proposed method can be extended to other procedures reporting type-III analyses, such as GENMOD and GLM.

INTRODUCTION

In medical, psychiatry, and social researches, missing data or incomplete data commonly occur due to various reasons, such as lost follow-up and nonresponses. Rubin (1976) classified missing mechanisms into three categories: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

Multiple Imputation was proposed by Rubin (1987) for making inferences on multiply imputed data, and has become the most popular method for analyzing missing data. Roughly speaking, missing values in the original data set are imputed independently M times to generate M completed data sets. Then a standard statistical procedure is applied to each complete dataset separately. Finally, all M analytical results are combined to form a single inference. Rubin and Schenker (1986) and Rubin (1987) developed a combining rule for parameter estimates from regression analyses. Li et al. (1992) proposed a method to combine p-values. SAS procedures MI and MIANALYZE are developed to implement these methods, which greatly facilitate the application of multiple imputations for researchers.

However, the current SAS procedure MIANALYZE is unable to summarize type-III analyses generated from SAS procedures, such as MIXED, GENMOD, and GLM, when they are executed on multiply imputed datasets. In the literature, there are different ways to calculate the sums of squares in order to compute F-test statistic and then p-value; as discussed in Goodnight (1978) and Herr (1986), there are at least 3 approaches, commonly called Type-I, II and III sums of squares. (This notation seems to have been introduced into the statistics world from the SAS package but is now widespread.) The type-III test tests for the presence of a main effect after adjusting for the other main effects and their interactions with the main effect being tested, and therefore we should consider this type in the presence of significant interactions.

In this manuscript, to make PROC MIANALYZE applicable for summarizing type-III analyses from multiple imputations, we create a macro named “type3_MI_mixed”, which can be applied with PROC MIXED. Based on this idea, we can create similar macros for other SAS procedure reporting type-III analyses, such as PROC GENMOD and PROC GLM.

METHODS

In order to combine type-III analyses obtained from multiple imputations, the main challenge is how to combine multiple test statistics which follow chi-squared distributions or F-distributions under null hypotheses. Fortunately, Raghunathan and Dong (2011) established a theoretical framework of combining random variables with F-distributions in the setting of ANOVA.

Suppose that, in an F-test statistic obtained from a complete dataset, where s_N is the numerator mean squares with expectation σ_N^2 and ν_N degrees of freedom and s_D is the denominator mean squares with expectation σ_D^2 and degrees of freedom ν_D. Under null hypothesis $\sigma_N^2 = \sigma_D^2$, the ratio $\nu_N s_N / \nu_D s_D$ is a pivotal statistic associated an F-distribution with degrees of freedom (ν_N, ν_D). Based on M imputed complete datasets, there are mean squares $s_N^{(l)}$ and $s_D^{(l)}$ associated with degrees of freedom $\nu_N^{(l)}$ and $\nu_D^{(l)}$, respectively, $l = 1, \dots, M$. Define $A_N = \sum_l 1/s_N^{(l)} /M$, $B_N =$

$\sum_l 1/(\nu_N^{(l)} \times s_N^{2(l)})/M$, and $C_N = \sum_l \left(1/s_N^{(l)} - A_N\right)^2/(M-1)$. Similarly, A_D, B_D, and C_D are defined for the denominator mean squares. By matching the posterior mean and variance, Raghunathan and Dong (2011) proposed to use $F_{MI} = A_D/A_N$ as the multiple-imputation adjusted F-statistic with the degrees of freedom (r_N, r_D), where $r_N = 2A_N^2/(2B_N + (M+1)C_N/M)$ and $r_D = 2A_D^2/(2B_D + (M+1)C_D/M)$.

Therefore, by invoking option "type3" in PROC MIXED, appropriate inferences can be drawn based on the above theory. Because the MIXED procedure allows for random effects, the resulting type-III ANOVA tables use different denominators in F-tests. To be specific, the column "Error Term" in the SAS output like Table 1 indicates the denominator used in F-tests. The developed macro can accommodate such differences.

This method can be extended to other procedures such as PROC GLM and PROC GENMOD. The type-III analyses from PROC GLM are based on F-tests, which are of a common denominator, the mean residual squares. Therefore, the method described earlier for PROC MIXED can be directly used for PROC GLM.

For GENMOD, the type-III analyses are a little different from those obtained by MIXED and GLM. In the output from GENMOD, the type-III analyses are based on likelihood ratio tests, which are chi-square tests instead of F-tests. It is straightforward to obtain corresponding formulas based on the above theory. To abuse the notation, assume $\Lambda = \nu_N s_N$ is a likelihood ratio statistic associated with a chi-squared distribution with degrees of freedom ν and note that actually $\sigma_N^2 = 1$. Therefore, we propose to use $\Lambda_{MI} = A_N^{-1}$ as the multiple-imputation adjusted chi-squared statistic with degrees of freedom r_N.

EAXMPLE

We use the dataset named Sp, from the Example 56.1 in SAS online documentation for the MIXED procedure. The original data set is complete, and we randomly delete two values of the response variable Y to mimic a dataset with missing values. In this dataset, covariates Block, A, and B are categorical variables with 4, 3 and 2 levels, respectively. Thus, the MIANALYZE procedure cannot provide proper inferences on the effects of Block and A.

```
data sp;
     input Block A B Y @@;
     datalines;
   1 1 1  56  1 1 2  41
   1 2 1  50  1 2 2  36
   1 3 1  .  1 3 2  35
   2 1 1  30  2 1 2  25
   2 2 1  36  2 2 2  28
   2 3 1  33  2 3 2  30
   3 1 1  32  3 1 2  24
   3 2 1  31  3 2 2  27
   3 3 1  15  3 3 2  19
   4 1 1  30  4 1 2  25
   4 2 1  35  4 2 2  .
   4 3 1  17  4 3 2  18
   ;
```

Firstly, the data setis imputed as follows:

```
proc mi data=sp seed=1305417 out=outsp;
     class Block A B;
     monotone reg (y);
     var Block A B y;
run;
```

Then we invoke the MIXED procedure by fitting the split-plot model with random block effects. Table 1 shows the type-III ANOVA table for the first imputed data set. Note that the F-tests of "A" and "Block" use a different denominator from others. Additionally, the data set containing the type-III table is generated by calling ODS, "ods output Type3=type3table".

```
proc mixed data=outsp method=type3;
     by _Imputation_;
     class A B Block;
     model Y = A B A*B;
     random Block A*Block;
     ods output Type3=type3table;
run;
```

Type 3 Analysis of Variance								
Source	**DF**	**Sum of Squares**	**Mean Square**	**Expected Mean Square**	**Error Term**	**Error DF**	**F Value**	**Pr > F**
A	2	180.122499	90.061249	Var(Residual) + 2 Var(A*Block) + Q(A,A*B)	MS(A*Block)	6	2.31	0.1799
B	1	295.259767	295.259767	Var(Residual) + Q(B,A*B)	MS(Residual)	9	13.84	0.0048
A*B	2	31.287816	15.643908	Var(Residual) + Q(A*B)	MS(Residual)	9	0.73	0.5070
Block	3	1639.884994	546.628331	Var(Residual) + 2 Var(A*Block) + 6 Var(Block)	MS(A*Block)	6	14.04	0.0040
A*Block	6	233.541676	38.923613	Var(Residual) + 2 Var(A*Block)	MS(Residual)	9	1.82	0.2004
Residual	9	192.056992	21.339666	Var(Residual)	.	.	.	.

Table 1: Type-III Tests Based on the First Imputed Data Set of Sp from PROC MIXED

To implement the proposed method, we create a macro named “type3_MI_mixed”, which can be used to summarize type-III analyses obtained from multiply imputed datasets. The proposed macro uses the dataset containing type-III table as the argument, for example “type3table” here. Then by invoking “%type3_MI_mixed(type3table);”, this macro automatically identifies denominators and numerators of F-tests and adjust them with multiple imputations. Table 2 shows the output summarized by macro “type3_MI_mixed”. The column “Source” shows all variable names, continuous or categorical; “# of imputation” shows the number of multiple imputations; “DF” shows the degrees of freedom of corresponding variables adjusted by the multiple imputation; “Error Term” indicates the term used as the denominator; “Error DF” shows multiple imputation adjusted degrees of freedom for the error term; “MI adjusted F” indicates the values of F-statistics adjusted by multiple imputations; “p_value” presents p-values adjusted by multiple imputation. It is clear the F-tests for A and Block use a different error term from others.

SUMMARY AND CONCLUSIONS

In this manuscript, we develop a method which provides multiple imputation adjusted type III analyses based on the MIXED procedure. The proposed method accommodates multi-level categorical variables and provided adjusted p-values for overall effects of corresponding variables instead of separate levels, which cannot be addressed by the MIANALYZE procedure. Furthermore, we can extend methods developed here to the GENMOD and GLM procedures. Due to different structures of type III tables, proper changes are made in corresponding macros. Only the macro for MIXED procedure is presented in Appendix. We are pleased to provide other two if readers are interested.

Obs	Source	# of imputation	DF	Error Term	Error DF	MI adjusted F	p-value
1	A	5	1.64189	MS(A*Block)	5.37281	2.9873	0.13682
2	A*B	5	1.23763	MS(Residual)	6.39841	1.6808	0.24724
3	A*Block	5	5.37281	MS(Residual)	6.39841	2.3861	0.15168
4	B	5	0.82907	MS(Residual)	6.39841	14.2281	0.00969
5	Block	5	2.56534	MS(A*Block)	5.37281	11.9395	0.00912

Table 2: Inference of Multiple Imputation Results of the Data Set Sp from PROC MIXED

REFERENCES

Goodnight, J. H. (1980). "Tests of hypotheses in fixed effects linear models." *Communications in Statistics - Theory and Methods* 9: 167-180.

Herr, D. G. (1986). "On the History of ANOVA in Unbalanced, Factorial Designs: The First 30 Years." *The American Statistician* 40: 265-270.

Raghunathan, T. E. and Dong, Q. (2011). "Analysis of variance from multiply imputed data sets." Unpublished manuscript. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan. Retrieved from http://www-personal.umich.edu/~teraghu/Raghunathan-Dong.pdf.

Rubin, D. B. (1976). "Inference and missing data." *Biometrika* 63: 581–590.

Rubin, D. B. and Schenker, N. (1986). "Multiple imputation for interval estimation from simple random samples with ignorable nonresponse." *Journal of the American Statistical Association* 81: 366-374.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

APPENDIX

```
%macro type3_MI_mixed(type3table);
*Identify Terms as Denomiator*;
data type3table;
*length Denominator $50;
set &type3table;
Denominator = trim(scan(ErrorTerm,2,'()'));
run;

proc freq data=type3table noprint;
where Denominator~=' ';
table Denominator /out=Denominator_name;
run;

*Identify Terms as Numerators*;
data error_set;
set type3table;
keep _imputation_ source df MS;
run;

proc sort data=error_set;
by source;
run;

data Denominator_name;
set Denominator_name;
rename Denominator=source;
run;
```

```
proc sort data=Denominator_name;
by source;
run;

*Merge Denominators to Each Error Source*;
data error_set;
merge error_set (in=in1) Denominator_name (in=in2);
by source;
if in2=1;
rename source=denominator DF=de_DF MS=MSE;
drop count percent;
run;

proc sort data=error_set;
by _imputation_ Denominator;
run;

proc sort data=type3table;
by _imputation_ Denominator;
run;

data type3table;
merge type3table error_set;
by _imputation_ Denominator;
label MSE='Mean Squared Error';
run;

*Adjust F-statistics by Multiple Imputation*;
data type3table;
set type3table;
An=1/MS;
Bn=1/(MS**2 * DF);

Ad=1/MSE;
Bd=1/(MSE**2 * de_DF);
run;

proc means data=type3table noprint;
class source;
output out=MIanalyze mean(An Bn Ad Bd)=ave_An ave_Bn ave_Ad ave_Bd var(An Ad)=ave_Cn
ave_Cd max(_Imputation_)=M;
run;

data MIanalyze;
set MIanalyze;
rn=2* ave_An**2 /( 2*ave_Bn+(M+1)*ave_Cn/M );
rd=2* ave_Ad**2 /( 2*ave_Bd+(M+1)*ave_Cd/M );
MI_F=ave_Ad/ave_An;
p_value=1-PROBF(MI_F,rn,rd);
run;

data finaloutput;
set MIanalyze;
where Source~=' ' and p_value~=.;
keep Source M rd rn MI_F p_value;
label M='# of Imputation’ rn=DF rd='Error DF' MI_F='MI adjusted F' p_value='p-value';
run;

proc sort data=finaloutput;
by source;
run;

proc sort data=type3table;
by source;
```

```
run;

data finaloutput;
merge finaloutput type3table (where=(_imputation_=1) keep=_imputation_ source
ErrorTerm);
by source;
run;

proc print data= finaloutput label;
where p_value~=.;
var source M rn ErrorTerm rd MI_F p_value;
run;

%mend;
```

CONTACT INFORMATION

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

Binhuan Wang
New York University School of Medicine, Department of Population Health
650 1st Ave Rm 578
New york, NY 10016
Phone: (202) 263-0016
Email: binhuan.wang@nyumc.org
Web: https://files.nyu.edu/wangb09/public/