运用 SAS 对多重填补数据集进行综合统计推断

——SAS 9 中的多重填补及其统计分析过程（二）

第二军医大学卫生统计学教研室(200433)　曹　阳　贺　佳

在 SAS 9 中，根据用户的设置的 m 值，MI 过程将对原来不完整的数据集填补 m 次（$m>1$），产生 m 个完整数据集，然后可以用任何针对完整数据集的标准 SAS 过程对它们进行分析。

设我们感兴趣的总体参数为 θ 和 σ^2，它们的点估计值分别是 θ 和 $\hat{\sigma}^2$，用相同的过程对每个填补数据集分别进行分析，会得到 $(\theta_1, \hat{\sigma}_1^2)$、$(\theta_2, \hat{\sigma}_2^2)$、…、$(\theta_m, \hat{\sigma}_m^2)$，结合这 m 组参数对 θ 的估计为：

$$\theta = \frac{1}{m}\sum_{i=1}^{m}\theta_i \tag{1}$$

考虑到填补后的数据的变异来自两个地方，一是填补数据集间的变异，二是填补数据集内的变异。因此方差的估计由两部分组成，一是填补内方差：

$$\sigma_w^2 = \frac{1}{m}\sum_{i=1}^{m}\hat{\sigma}_i^2 \tag{2}$$

二是填补间方差：

$$\sigma_B^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\theta_i - \theta)^2 \tag{3}$$

总体参数的方差估计 σ_T^2 为：

$$\sigma_T^2 = \sigma_W^2 + (1+\frac{1}{m})\sigma_B^2 \tag{4}$$

σ_T^2 的平方根就是 θ 的总的标准误。可以看出，当没有缺失数据时，θ_1、θ_2、…、θ_m 都是一样的，σ_B^2 等于 0，σ_T^2 等于 σ_W^2。因此，从方差的角度来说，σ_B^2 的大小反映了缺失数据与观察到的数据相比相对包含了多少信息。

θ 近似的 95% 可信区间估计是 $\theta \pm 2\sqrt{\sigma_T^2}$，但是计算 θ 的可信区间的更好方法是用近似的 t 分布：

$$\theta \pm t_v \sqrt{\sigma_T^2} \tag{5}$$

公式 5 中的自由度 v 的计算公式为：

$$v = (m-1)\left(1+\frac{m\sigma_W^2}{(m+1)\sigma_B^2}\right)^2 \tag{6}$$

对总体参数 θ 的缺失部分信息的估计是：

$$\gamma = \frac{r+2/(v+3)}{r+1} \tag{7}$$

公式 7 中的 r 是由于数据缺失造成的方差的相对增量，其计算公式为：

$$r = \frac{(1+m^{-1})\sigma_B^2}{\sigma_W^2} \tag{8}$$

γ 与 r 都是很有用的诊断指标，揭示了 θ 的估计在多大程度上受到了数据缺失的影响。应该注意的是，当填补次数 m 趋向无穷大时，总方差 σ_T^2 等于 σ_W^2 与 σ_B^2 之和，θ 的可信区间的估计是基于正态分布（$v=\infty$）。自由度 v 同时受到填补次数和 σ_W^2 与 σ_B^2 的相对大小的影响，当 σ_B^2 远远大于 σ_W^2，自由度趋向于最小值 $m-1$，当 σ_W^2 远远大于 σ_B^2 时，自由度趋向于无穷大。如果计算所得的 v 比较小时，如小于 10，建议增加填补的次数以获得更高的效率；然而当自由度较大时，增加填补次数的意义不大。从方差的角度来说，多重填补的效率大约为 $\left(1+\frac{\gamma}{m}\right)^{-1}$[1-3]。

MIANALYZE 的语句说明

MIANALYZE 过程的语句构成及语法如下[4]：

PROC MIANALYZE <选择项>;

BY 变量名或变量列表;

CLASS 变量名或变量名列表;

MODELEFFECTS 效应变量或效应变量列表;

<标签:>TEST 公式 1 <,…,<公式 k>><</选择项>;

STDERR 变量名或变量列表;

PROC MI 语句中包含的选择项较多，主要分为三类，第一类是对输入数据集进行定义，有 DATA=、PARMS=、PARMINFO=、COVB= 和 XPXI；第二类是定义统计量，有 THETA0=、ALPHA= 和 EDF=；第三类是定义显示的输出结果，有 WCOV、BCOV、TCOV 和 MULT。下面主要的选择项加以介绍。

DATA= 数据集，该选择项定义了输入数据集。如果输入的是特定结构的数据集，则其中必须有一个 TYPE 变量表示该数据集包括了填补数据集的哪些估计值。当 TYPE=EST 时，表示数据集包括了参数估计值和协方差矩阵；TYPE=COC 表示数据集包括的是样本均数、样本含量、协方差矩阵；TYPE=CORR 表示数据集包括的是样本均数、样本含量、标准误和相关系数矩阵。如果输入数据集不是特定结构的数据集，该数据集中所包含参数估计值的变量和对应的标准误的变量分别由 MODELEFFECTS 和 STDERR 语

句说明。

PARMS <(CLASSVAR=分类变量的类型)>=数据集,该选择项定义了根据填补数据集计算得到的参数估计值。如果没有使用 COVB=数据集选择项,则 PARMS 所定义的数据集中还包括了参数估计值所对应的标准误。如果在用 CLASS 语句定义了分类变量,还可以在 PARMS 后跟上 CLASSVAR=分类变量的类型这一选择项定义读取分类变量水平的方式。

COVB=数据集,该选择项定义了根据填补数据集计算得到的参数估计值的协方差矩阵。如果使用这一选择项,必须使用 PARMS=数据集这一选择项。

XPXI=数据集,该选择项定义了根据填补数据集计算得到的参数估计值的$(X'X)^{-1}$矩阵。PROC MIANALYZE 可根据从 PARMS=数据集中读取到的标准误和$(X'X)^{-1}$计算协方差矩阵。

THETA0|MU0=数值,该选择项定义了对效应变量进行 t 检验时,无效假设 H_0:$\theta=\theta_0$ 中 θ_0 的值。如果只定义了一个 θ_0 值,则对所有的效应变量都按这个值进行 t 检验。如果定义了多个 θ_0 值,则这些值与此同时 MODELEFFECTS 语句中定义的效应变量的顺序相一致。对于 CLASS 语句定义的分类效应变量,不进行检验。

ALPHA=p 值,该选择项定义了估计参数 $100(1-p)\%$可信限时的 p 值。

EDF=数值,该选择项定义了完整数据集的自度度,用于计算每一个参数估计中的校正自由度。默认值为∞,不对自由度进行校正。

MNLT|MULTIVARIATE 选择项要求对参数进行多元统计推断,采用的是单变量统计推断扩展出来的 Wald 检验。

BCOV、WCOV 和 TCOV 这三个选择项分别要求在结果中显示填补间协方差矩阵、填补内协方差矩阵和总的协方差矩阵。

BY 语句指明了分组变量,MIANALYZE 过程根据这一变量将数据集分成若干组分别进行统计分析。在使用 BY 语句时,数据集必须已经根据 BY 语句中的变量排过序。如果数据集没有按照 BY 语句中的变量进行过升序排序,可以使用下面的方法:

一是用 SORT 过程按 BY 语句中相同的变量进行排序;

二是在 BY 语句中使用 NOTSORTED 或 DESCENDING 选择项,NOTSORTED 选择项并非是指数据集没有排过序,而是指只将数据按 BY 中指明的变量的数值分组,但是不必将这些组别按照数值大小或是字母顺序排序;

三是用 DATASETS 过程对 BY 语句中的变量创建一个索引。

CLASS 语句是 SAS 9 中新增添的语句,它定义了 MODELEFFECTS 语句中的哪一个变量是分类变量,这一变量可以是数字型也可以是字符型。分类水平是根据 CLASS 语句所指明的变量的格式化取值所决定的,可参考 SAS/BASE 中的 FORMAT 过程。

MODELEFFECTS 语句中列出了需要进行分析的效应,这些效应可以是一个变量也可以是变量组合,用变量名和特定运算符表示。这些变量可以是分类变量(由 CLASS 语句说明)或是连续型变量。没有在 CLASS 语句中说明的变量被认为是连续型变量。可以使用交互和嵌套运算符产生交互效应和嵌套效应。通常的定义形式为:

```
CLASS A B C;
MODELEFFECTS X1 X2 X1*X2*A*(B C);
```

其中 A、B 和 C 是分类变量,$X1$ 和 $X2$ 为连续型变量。

STDERR 语句列出了 MODELEFFECTS 语句中的效应变量的标准误,这时参数估计值和标准误同时被作为变量保存在 DATA=数据集选择项所指明的数据集中。如果通过 DATA=数据集定义的数据集不是特定结构的 SAS 数据集,则必须用 STDERR 语句指明数据集中的标准误变量或参数估计值变量。

只有在 MODELEFFECTS 语句中的每一个效应变量都是连续型变量时,才能使用 STDERR 语句。STDERR 语句中定义的标准误顺序是与 MODELEFFECTS 中所对应的效应变量的顺序是一一对应的。例如,下面一段程序定义了所要分析的效应及需要用到的各效应对应的标准误:

```
PROC MIANALYZE;
    MODELEFFECTS Y1-Y3;
    STDERR SY1-SY3;
RUN;
```

TEST 语句是 SAS 9 中新增添的语句,它对关于参数 β 的线性假设进行检验。在同一个 TEST 语句中,通过一个 F 检验对一个或多个无效假设(H_0:$L\beta=c$)进行检验。

该语句中的每一个公式定义了一个线性假设,其中 L 是线性假设的系数矩阵,c 是一个常数向量。假设我们的总体参数 θ 的点估计和协方差估计分别为 θ 和 Σ,θ_i 和 Σ_i 是第 i 个填补数据集的估计值。则对于一个给定的 L,在第 i 个填补数据集中,线性函数 $L\theta$ 的点估计和协方差估计分别为:$L\theta_i$ 和 $L\Sigma_iL'$。对每一个 TEST 语句,程序的运行结果中会显示出每一个线性成分的合并估计值和标准误。

在 MIANALYZE 过程中可以使用多个 TEST 语句,通过在各个 TEST 语句前加上标签加以区分。如

果未加标签，则在运行结果中，SAS 会自动以“Test j”表示第 j 个 TEST 语句的检验结果。每一个 TEST 语句中可以有多个等式，每个等式间用逗号分开。假设我们的数据集中有 $X1$、$X2$、$X3$ 和 $X4$ 这 4 个变量，我们可以编写如下的程序：

```
PROC MIANALYZE;
    VAR X1 X2 X3 X4;
    test1: TEST X1+X2=0;
    test2: TEST X1+X2;
    test3: TEST X2=X3=X4;
    test4: TEST X2=X3, X3=X4;
RUN;
```

若在等式中没有等号出现时，SAS 默认该表达式的值为 0，即上面的程序中 test1 与 test2 是等价的。

实　例

我们结合上一篇文章中的数据集，对 SAS 中的 MIANALYZE 过程加以应用。

我们上次已经运用 PROC MI 过程生成了一个填补数据集 outExp，该数据集中的变量 –Imputation– 表明了是哪一次填补。

首先，我们可以用 PROC UNIVARIATE 过程计算每一次填补数据的样本均数和标准误，程序如下：

```
proc univariate data=outExp noprint;
    var Oxygen Time Rate;
    output out=outuni mean=Oxygen Time Rate
    stderr=SOxygen STime SRate;
    by –Imputation–;
run;
```

下面的程序是利用 PROC UNIVARIATE 过程的结果进行综合统计推断：

```
proc mianalyze data=outuni edf=30;
    modeleffects Oxygen Time Rate;
    stderr SOxygen STime SRate;
run;
```

因为原数据集中共有 31 个观测，我们在上面的程序中指定完整数据集的自由度为 30。SAS 9 的输出结果分为两部分，一部分是每一个单变量统计推断的填补间方差、填补内方差和总方差；一部分是各变量的总体均数的估计值、95%可信区间以及各变量的总体均数是否等于 0 的 t 检验结果。

对于可以将参数估计值和协方差矩阵输出到一个数据集中的过程（如 PROC REG），可以直接用 PROC MIANALYZE 调用其结果。例如，我们想对氧气摄入量和时间、心率之间进行回归分析，并对回归方程中的各个参数进行综合统计推断，其程序如下：

```
proc reg data=outExp outest=outreg covout noprint;
    model Oxygen= Time Rate;
    by –Imputation–;
proc mianalyze data=outreg;
    modeleffects Intercept Time Rate;
run;
```

上面的程序会给出结合 5 个填补数据集对回归方程的截矩、时间的回归系数和心率的回归系数的综合统计推断结果，包括总体参数的估计值、95%可信区间及总体参数是否等于 0 的 t 检验的结果。

在 MI 方法出现以前，简单删除和单重填补（single imputation）是处理缺失值问题的主要方法，它们难以保证对总体参数进行有效的统计推断和对数据的利用效率。而 MI 方法和相应的统计软件的出现，弥补了这些缺陷。虽然 MI 并不是处理缺失数据的唯一方法，也不一定是最好的方法，但在实际应用中，涉及大量参数估计的探索性分析或是多目标分析时，消除缺失数据的干扰可能是我们关心的主要问题，简便易用的近似解决方案比特定的、难于使用的方法更受欢迎，这时 MI 就能充分发挥它的优点。希望统计人员们能够借助 SAS 9 等统计软件在处理缺失数据方面的最新发展，提高数据的利用效率和统计推断的质量。

参 考 文 献

1. Rubin DB. Inference and missing data. Biometrika, 1976, 63(3): 581-592.
2. Rubin DB. Multiple imputation: a primer. Statistical Methods in Medical Research, 1999, 8(1): 3-15.
3. James MR. Inference for imputation estimators. Biometrika, 2000, 87(1): 113-124.
4. SAS Institute Inc. SAS/STAT 9 User's Guide. North Carolina: SAS Institute Inc, 2003.