·计算机应用·

运用 SAS 对不完整数据集进行多重填补
——SAS 9 中的多重填补及其统计分析过程（一）

第二军医大学卫生统计学教研室(200433)　曹　阳　张罗漫

在所有的实验研究和调查研究领域，数据缺失是一个普遍存在的问题。缺失数据会增加分析任务的复杂性、造成结果偏倚、降低统计工作的效率。尤其是在完全观测与不完全观测之间存在系统差异的情况下，运用常规统计分析方法对不完整数据集所做出的结果，不能代表整体。在近二、三十年来，多重填补(multiple imputation，MI)方法被认为是解决这一问题的首选方法，该方法由 Donald B. Rubin 在 20 世纪 70 年代首先提出[1,2]。与通常用平均值代替缺失值或其他简单填补(simple imputation)方法的不同之处在于，MI 方法对每一个缺失值用一套可能的值进行填补，以反映缺失值的不确定性，从而产生若干个完整数据集；然后，用针对完整数据集的统计方法对每一个填补数据集分别进行统计分析，把得到的结果进行综合，进而产生最终的统计推断(图 1)。这种方法能够反映出由于数据缺失造成的统计推断结果的不确定性[3]。随着计算方法的不断成熟和相应统计软件(如 Amelia、Solas、Norm、Iveware 和 Emcov 等)的出现，该方法已被越来越多地应用于生物医学、行为科学和社会科学领域。

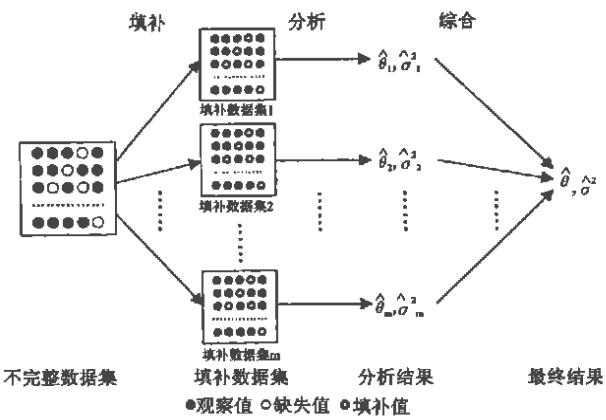

图 1　多重填补步骤及其统计推断原理

在 SAS/STAT 软件中，从 8.0 版本中开始引入对缺失数据进行多重填补及其统计分析的 MI 和 MIANALYZE 两个过程，并在 8.1 和 8.2 版本中对这两个过程的功能做了进一步修订。MI 过程用于对含有 p 个变量的不完全数据集产生 m 个填补数据集(m 值由用户定义)。它所使用的方法结合了这 m 个数据集中的变异性。在对每一个填补数据集用 SAS/STAT 中的标准过程进行分析之后，就可以用 MIANALYZE 过程进行综合统计推断。MI 过程中提供了 3 种方法对缺失值进行填补。对于单调缺失(monotone missing)模式，可使用基于多元正态性假设的参数回归方法或采用趋势得分(propensity score)的非参数方法；对于任意缺失(arbitrary missing) 模式，可使用基于多元正态性假设的马尔科夫链蒙特卡罗(Markov chain Monte Carlo，MCMC)方法[4,5]。

在 SAS 8 中，MI 和 MIANALYZE 过程还只是两个实验过程，所以在 SAS 的用户参考手册中并没有对其进行介绍。在 SAS 9 中，MI 和 MIANALYZE 已经成为 STAT 模块中的两个正式过程。和 SAS 8 相比，SAS 9 的 MI 过程中，对 MCMC 和 MONOTONE 语句新增加了 REGPMM 选择项，可以用预测均数匹配(predictive mean matching)法把一个缺失值用接近预测值的数值进行填补。此外，还增加了 CLASS 语句，对分类变量进行定义。在 MONOTONE 语句中，新增加了 LOGISTIC 和 DISCRIM 选择项，分别用 Logistic 和判别方法填补缺失值。DATA＝选择项的功能也得到了增强，在输出数据集时可以同时包括每一个观测的参数估计值和相应的标准误。SAS 9 的 MIANALYZE 过程中，新增加了 TEST 和 CLASS 语句。下面，对这两个过程分别加以介绍[6]。

MI 过程简介

MI 过程的语句构成及语法如下：

PROC MI ＜选择项＞；
　　BY 变量名或变量列表；
　　CLASS 变量名或变量名列表；
　　EM ＜选择项＞；
　　FREQ 变量名；
　　MCMC ＜选择项＞；
　　MONOTONE ＜选择项＞；
　　TRANSFORM 变换方法（变量名或变量列表＜/选择项＞）＜... 变换方法（变量名或变量列表＜/选择项＞）＞；

VAR 变量名或变量列表;

BY 语句指明了分组变量,MI 过程根据这一变量将数据集分成若干组分别进行多重填补。

CLASS 语句是 SAS 9 中新增添的语句,它定义了 VAR 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。

EM 语句是在假设数据集服从多元正态分布的基础上,根据 EM (expectation and maximization)法则计算含有缺失值数据集的极大似然估计。

FREQ 语句指明了频数变量,该变量表示了数据集中其他变量的每一个观测值出现的次数。

MCMC 语句指明对于任意缺失模式的数据集,基于数据服从多元正态性假设,采用 MCMC 方法进行填补。

MONOTONE 语句用于对单调缺失模式数据集中的连续型变量或 CLASS 语句中指明的分类变量进行填补,既可以使用参数的回归方法,也可以用基于趋势得分的非参数方法。MCMC 语句和 MONOTONE 语句不能同时使用。如果程序中没有使用这两条语句,默认使用的是 MCMC 方法。

TRANSFORM 语句指明了在填补过程之前进行变量变换的变量。在填补数据集之前,填补值被转换回原来的数量单位。

VAR 语句指明了要进行分析的变量。如果省略了 VAR 语句,则对所有的数值型变量都进行分析。

在 MI 过程中,可以只使用一句 PROC MI 语句。

MI 过程中主要选择项的说明

1.PROC MI 语句中的选择项

(1)ALPHA=α,指明在进行均数的区间估计时,其置信水平为$(1-\alpha)\times 100\%$,$0<\alpha<1$,缺省值为 $\alpha=0.05$。

(2)DATA=数据集名称,指明 MI 过程进行分析的数据集,缺省状态下,使用最近一次创建的数据集。

(3)MAXIMUM=数值 1<数值 2…>,指明对变量进行填补时,填补值的上限,如果将要填补的值超了这一上限,MI 过程将重新抽取一个填补值。如果只定义了一个数值,则所有变量的填补值的上限都由这一数值决定。如果定义了一个以上的数值,则该选择项必须配合 VAR 语句一起使用,数值的顺序与 VAR 语句中的变量名一一对应。缺失值"."表示对应的变量的填补值没有限制。缺省状态下是"MAXIMUM=.",即对所有变量的填补值没有上限限制。MAXIMUM=选择项与 MINIMUM=、ROUND=选择项相关联,这些选择项可以让填补值和观察到的变量值更趋于一致。只有在使用单调回归方法或没有 PMM 选择项的 MCMC 方法时才能使用这些选择项。

如果只对第一个变量定义了填补值的上限,则必须在这个值后面定义一个缺失值,否则 MI 过程会对所有的变量都使用这一限制。例如"MAXIMUM=100 ."表示只对第一个变量设置最大填补值为 100 的限定,而对后面的所有变量的填补值没有限定。"MAXIMUM=. 100"表示只对第二个变量设置最大填补值为 100 的限定,而对其他变量的填补值没有限定。

(4)MINIMUM=数值 1<数值 2…>,指明对变量进行填补时,填补值的下限,其使用说明同 MAXIMUM=。

(5)NIMPUTE=数值,定义填补的次数,默认为 5 次。可设置 NIMPUTE=0,不对数据集进行填补,而只显示数据缺失模式、简单的描述性统计量和期望最大化估计值。

(6)OUT=数据集名称,创建经过填补后的数据集。在这一数据集中增加了一个索引变量"–Imputation–",用于指明是第几次填补。在每一次填补中,原始数据集中的缺失值都被填补值替换。

(7)SEED=数值,设定一个正整数值,MI 过程使用这一数值作为伪随机数的种子。缺省状态的取值是通过计算机当时的时间计算出的数值。如果为了在同样的条件下重复结果,必须在每次分析时使用同样的种子,而不能依赖于计算机的时间。

(8)SIMPLE,显示简单的单变量描述性统计量和通过可利用的观测计算得到的变量间的两两相关系数。

(9)SINGULAR=p,这是 SAS 9 中新增添的选择项,它确定了标准化变量的协方差矩阵奇异性评判标准,默认值是 SINGULAR=1E−8。

2.EM 语句中的选择项

EM 法则是一种用于不完整数据集参数模型中极大似然估计的技术。EM 语句要求对一个含有缺失值的数据集,在多元正态分布假设基础上,计算均数和协方差矩阵的极大似然估计。MI 过程用可利用观测或完整观测的均数和标准差作为 EM 法则的初始估计值。

在 EM 语句中可以使用下面的选择项:

(1)CONVERGE=p,设置收敛标准,$0<p<1$。当各次迭代间的参数估计值的变化小于 p 时,可以认为迭代是收敛的。当参数的绝对值大于 0.01 时,变化是指迭代间的相对变化,否则指的是绝对变化。缺省状态下,p 值为 0.0001。

(2)MAXITER=正整数值,指定 EM 法则的最大迭代次数,缺省值是 MAXITER=200。

3.MCMC 语句中的选择项

MCMC 语句定义了运用 MCMC 方法进行填补时

的具体内容，可以使用的主要选择项如下：

(1)CHAIN＝SINGLE|MULTIPLE，指明是对所有的填补使用同一条链，还是每次填补使用单独的链。缺省值是 CHAIN＝SINGLE。

(2)IMPUTE＝FULL|MONOTONE，指明是对所有的缺失数据进行填补，还是只对部分缺失数据进行填补以把数据集转换成单调缺失模式。缺省值是 IMPUTE＝FULL。当指明了 IMPUTE＝MONOTONE 时，按照 VAR 语句中列出的变量顺序把数据集填补为单调缺失模式。

(3)INITIAL＝EM ＜(选择项)＞，指明 MCMC 过程中初始的均数和协方差估计值。缺省值为 INITIAL＝EM，即使用 EM 的估计值作为 MCMC 过程的初始值。

(4)OUTEST ＝数据集名称，创建一个 TYPE＝EST 的数据集，用于存放每一次填补后的参数估计值。在这个数据集中有一个–Imputation–变量，用于表示是哪一次填补的结果。

(5)OUTITER ＜(选择项)＞ ＝数据集名称，创建一个 TYPE＝COV 的数据集，其中存放了填补步中每一次迭代使用的参数。数据集中用–Imputation–和–Iteration–分别表示填补的次数和迭代的次数。可使用的选择项有 MEAN、STD、COV、LR、LR–POST 和 WLF，分别表示均数、标准差、协方差、－2logLR 值、后验模型的－2logLR 值以及最差线性函数值。如果没有对选择项进行定义，数据集中只存放每次迭代所使用的均数。

4.MONOTONE 语句中的选择项

MONOTONE 语句中指明了对单调缺失的数据集的填补方法。在使用这个语句时，必须使用 VAR 语句，而且被填补数据集中呈现单调缺失模式的变量的顺序必须与 VAR 语句中所列出的变量顺序相一致。当同时使用 MCMC 语句和 MONOTONE 时，MI 过程不执行 MCMC 语句。MONOTONE 语句中可使用的选择项如下：

(1)DISCRIM ＜(被填补变量 ＜＝ 效应变量＞＜/ 选择项＞) ＞，用判别函数方法填补分类变量。要与 CLASS 语句一起使用。

(2)LOGITIC ＜(被填补变量 ＜＝ 效应变量＞＜/ 选择项＞) ＞，用 Logistic 回归方法填补分类变量。要与 CLASS 语句一起使用。

(3)REG|REGSSION ＜(被填补变量 ＜＝ 效应变量 ＞＜/ DETAILS＞) ＞，用回归方法填补连续变量。DETAILS 选择项用于显示每一次填补中回归模型的回归系数。当使用回归方法时，可以在 ROC MI 语句中使用 MAXIMUM ＝、MINIMUM ＝和 ROUND＝等选择项，以使填补值与观测值更加一致。

(4)REGPMM|REGPREMEANMATCH ＜(被填补变量 ＜＝ 效应变量 ＞＜/ DETAILS＞) ＞，用预测均数匹配法填补连续变量。

(5)PROPENSITY ＜(被填补变量 ＜＝ 效应变量＞＜/ 选择项＞) ＞，用趋势得分法填补连续型变量或分类变量。

5.TRANSFORM 语句

TRANSFORM 语句定义了数据变换的方法和进行变换的变量。在 MI 过程使用回归方法和 MCMC 方法时，假设数据服从多元正态分布。有时数据集中有些变量明显不服从正态分布，这时应该把它们转换成服从多元正态分布。使用了 TRANSFORM 语句之后，指定的变量在填补前进行了变换，过程中显示的所有结果都是变换后的数值。如果定义了 OUT＝选择项，则相应的变量被反变换后再创建填补后的数据集。

可用使用的变量变换方法有：

(1)BOXCOX，Box-Cox 变换，原变量 Y 被转换为 $\frac{(Y+c)^{\lambda}-1}{\lambda}$，其中 c 是一个常数，使得 $Y+c>0$，λ 是个大于 0 的常数；

(2)EXP，指数变换，原变量 Y 被转换为 $e^{(Y+c)}$，c 是常数；

(3)LOG，对数变换，原变量 Y 被转换为 $\log(Y+c)$，其中 c 是常数，使得 $Y+c>0$；

(4)LOGIT，logit 变换，原变量 Y 被转换为 $\log(\frac{Y/c}{1-Y/c})$，c 是大于 0 的常数，并且 $0<Y/c<1$；

(5)POWER，幂变换，原变量 Y 被转换为 $(Y+c)^{\lambda}$，其中 c 是常数，使得 $Y+c>0$，λ 是不等于 0 的常数。

涉及 c 和 λ 的地方可用 C ＝ 数值和 LAMBDA ＝ 数值这两个选择项。

实　例

结合下面的数据集，对 SAS 中的 MI 过程加以应用。

```
*--------一个关于健康状况的数据集------*
|本数据集是根据一些男性健康状况的数据改编的虚拟数据集，|
|数据集中的变量分别为：                  |
| Oxygen (氧气摄入量，ml/公斤体重/分钟)，|
| Time (跑完 2.4 公里所花费的时间，单位为分钟)，|
| Rate (奔跑时的心率)。                  |
*--------------------------------*;
data Example;
input Oxygen  Time Rate @@;
datalines;
43.509  10.27  175  44.313  10.14  182
```

(下转第 63 页)

足病历在存储、传输、操作等方面需要的，特别是具有数据库DBA(DataBase Administration)权限的用户可以任意察看、篡改病历内容。一种可行的解决办法是：对病历段的内容、签名、日期等关键字段进行加密和保护，使没被授权的用户看不懂病历——实现保密，恶意篡改会破坏密文甚至使之不能正确脱密还原——留下痕迹，从而实现了保密和防篡改。

病历的加/脱密算法应该保密性强、速度快，其使用对合法用户而言是透明的、觉察不到的。

5. 病历检索对数字病历的要求

在病历(病案)检索方面，病历检索人员不关心病历被修改了几次、做了什么修改，而关心修改后的终稿。这就要求电子病历系统能够在计算机内部对病历信息进行“组装”：去掉被删除的部分，插入(含增加)修改时增加的部分，替换掉被替换的部分，生成用于检索的内容。文本具有比图片、特殊标记高得多的检索价值。

另外，电子病历的开发还要关注到互联网的普及，以便支持医师在互联网上书写病历；不能指望每位医师都是打字高手，应该尽可能减少格式性、重复性、关联性内容的输入工作量，使医疗质量和工作效率都得到提高；电子病历的使用也要得到相关法规和部门的支持和认可。

电子病历的研究开发已经得到了国际范围的广泛关注，一些国家把它作为国家形象来抓，如美国、英国、日本、荷兰等地区电子病历已有了相当程度的研究和应用。国际上，公认电子病历应当具有三个内涵[4]：第一具有信息共享系统：医院的各个部门、科室在任何地方、任何时候都能调阅到病人所在医院的全部病历记录；第二具有预警系统：药物配制禁忌、医疗方法不正当的提示，是医疗的智能化；第三医疗信息资料库支持：内有电子图书、电子杂志以及关于病例治疗最新方法。在国内，电子病历开发和应用虽然频见媒体，但还没有真正意义上的电子病历，尚处起步阶段。电子病历是一个涉及医学、医疗设备、法规、加/脱密技术、计算机和网络技术等多领域的综合性大项目，综合性强、技术性高，需要政府、企业、科技人员等广泛参与，摆脱低层次重复，科学规划，分步实施。

参 考 文 献

1. 卫生部和国家中医药管理局. 医疗机构病历管理规定. http://www.moh.gov.cn/yzgl/index.htm.
2. 中国人民解放军总后勤部卫生部. 医疗护理技术操作常规. 第4版. 人民军医出版社，1998，8.
3. 刘志文，吴一民. 基于XML标准的电子病历实现技术. 微型机与应用，2001，(5)：37-39.
4. 北京尚无电子病历 只是病历电脑化管理. http://health.sohu.com/11/30/harticle16883011.shtml.

(上接第58页)

```
55.018  8.73  163  59.571  .      .
48.982  9.34  .    45.823  12.14  175
.       12.04 177  .       11.02  .
39.501  12.97 175  60.247  9.01   171
50.724  .     .    37.413  13.92  183
45.011  11.34 177  46.924  .      .
52.145  10.48 167  48.877  9.04   178
41.241  11.02 171  46.721  10.13  .
46.835  10.37 .    50.117  9.89   165
40.216  12.71 175  46.136  11.34  158
45.532  9.71  165  .       9.17   .
45.357  11.19 .    40.014  13.24  167
46.121  10.68 187  50.623  9.89   151
48.714  9.52  186  48.073  11.65  171
47.542  10.61 172
;
```

假设数据集中的数据服从多元正态分布，而且数据缺失模式为任意缺失。运用MI过程对数据集Example进行多重填补的最基本的程序如下：

```
proc mi data=Example seed=1000 out=outExp;
    var Oxygen Time Rate;
run;
```

程序运行结果略。原来的数据集被填补了5次，输出到名为outExp的数据集中。

参 考 文 献

1. Rubin DB. Multiple imputation: a primer. Statistical Methods in Medical Research, 1999, 8(1): 3-15.
2. Rubin DB. Inference and missing data. Biometrika, 1976, 63(3): 581-592.
3. James MR. Inference for imputation estimators. Biometrika, 2000, 87(1): 113-124.
4. Barnard J, Rubin DB. Small-sample degrees of freedom with multiple imputation. Biometrika, 1999, 86(4): 948-955.
5. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika, 1983, 70(1): 41-55.
6. SAS Institute Inc. SAS/STAT 9 User's Guide. North Carolina: SAS Institute Inc, 2003.