任意缺失模式缺失数据不同填补方法效果比较*

张　桥[1]　李　宁[2]　张秋菊[1]　刘美娜[1△]

【提　要】 目的　探讨任意缺失模式下缺失数据的填补方法，并对不同方法填补效果进行比较和评价。方法　结合我国北方绝经期妇女钙需要和膳食评估应用研究课题的数据，调用 SAS 软件中 IML 模块产生任意缺失模式模拟数据，通过 MI 和 MIANALYZE 过程实现缺失数据的填补，同时应用准确度和稳定度两个评价指标来评价各方法填补的效果。结果　PS 方法填补 3 次在本文模拟的任意缺失模式的缺失数据中填补效果最佳，MCMC 方法填补效果并不理想。结论　在填补任意缺失模式的缺失数据时，MCMC 并不是唯一的多重填补方法，通过多重填补的 PS 方法、PMM 方法和 REG 方法把数据填补成单调缺失后，再用相同方法进行一次填补也是一种可选择的填补方法。

【关键词】 缺失数据　任意缺失模式　多重填补　数据模拟

数据缺失是实验研究和调查研究中一个普遍存在的问题[1]，如何正确的处理、分析所缺失的数据在数据分析中占有重要地位。缺失数据的类型按照不同的分类方法可划分不同类别，按缺失机制分类和按缺失模式分类两种划分方法[2-3]。

按照由 Little 和 Rubin 在 1976 年提出的缺失机制分类，缺失数据可以分为完全随机缺失(missing completely at random，MCAR)、随机缺失(missing at random，MAR)和非随机缺失(not missing at random，NMAR)三类[4]。如果所缺失的数据发生的概率既与已观察到的数据无关也与未观察到的数据无关，则该缺失数据类型为 MCAR；如果缺失数据的发生概率与所观察到的变量是有关的，而与未观察到的数据特征无关，则该缺失数据类型为 MAR；若数据既不属于完全随机缺失也不属于随机缺失，那么该缺失数据类型就属于 NMAR[5]。按照数据缺失模式可以分为单调缺失模式和任意缺失模式两类[6-7]，为了简单明了可以通过图 1 来形象的理解，其中是 5 个变量，1～5 是 5 个样本，“×”表示数据能观察到，“.”表示数据缺失。

图 1　数据缺失模式

单调缺失模式如图 1(a)所示，对数据集进行适当的行列变换后，可以得到这样一个矩阵，它呈现出一种层级缺失的模式，矩阵中的元素 y_j 缺失时，则对任意的 $P \geqslant j$，元素 y_p 也是缺失的；任意缺失模式如图 1(b)所示，数据缺失具有随意性，没有任何规律可循，即使通过行列变换也无法看出任何规律。

对于任意缺失模式的数据处理，查阅相关文献发现常用的就是把缺失值直接删除即 Ad Hoc 法或多重填补(multiple imputation，MI)中的马尔科夫链蒙特卡洛(markov chain monte carlo，MCMC)方法[8]，对于纵向数据有时也采用单一填补中的 LOCF(last observation carried forward)方法[9]。本文将探讨 Ad Hoc 法、LOCF 填补、多重填补中的回归方法、预测均数匹配(predictive mean matching，PMM)方法、趋势得分(propensity score，PS)方法、MCMC 方法这六种方法对任意缺失模式下缺失数据的填补效果。

资料与方法

1. 资料来源

本文所用数据来源于国家科技支撑计划项目：我

* 基金项目：国家科技支撑计划(2011BAI09B02)
1. 哈尔滨医科大学公共卫生学院卫生统计学教研室(150081)
2. 宁波市疾病预防控制中心免疫预防所
△通信作者：刘美娜，E-mail：liumeina369@163.com

国北方绝经期妇女钙需要和膳食评估应用研究。此课题是一个为期两年人群干预研究，研究对象 282 名，通过分层随机方法分为四组，3 个钙干预组和 1 个信息干预组。分别在干预前、干预 1 年后、干预 2 年后三个时间点对干预对象进行调查和样品采集，获得研究对象的体格检查、一般情况、饮食情况、体力活动情况和心理与应对等信息，同时对研究对象进行骨密度检测，所采用仪器是美国 Norland XR-36 双能 X 线骨密度仪，包括腰椎、髋骨和全身骨三个部位，获得相应部位的骨密度 T 值。本文主要选用志愿者的身高、体重、年龄以及三次骨密度检查的腰椎骨密度 T 值作为模拟实验的参考数据。

2. 数据基本状况

参考数据中身高、体重、年龄和第一次腰椎骨密度 T 值为完整数据，共 282 例，第二次和第三次腰椎骨密度 T 值分别缺失 63 人和 80 人，因此剩余人数分别是 219 和 202 例。参考数据中各变量的均数和标准差见表 1。

表 1　参考数据各变量的均数和标准

	腰椎 T 值 1	腰椎 T 值 2	腰椎 T 值 3	身高	体重	年龄
均数	-1.85	-1.59	-1.82	158.06	61.84	57.98
标准差	2.08	2.11	2.06	5.30	8.40	3.86

表 2 是参考数据中各变量间的相关系数矩阵。

表 2　参考数据各变量的相关系数矩阵

	腰椎 T 值 2	腰椎 T 值 3	身高	体重	年龄
腰椎 T 值 1	0.94	0.92	0.32	0.33	-0.27
腰椎 T 值 2		0.94	0.30	0.33	-0.26
腰椎 T 值 3			0.27	0.31	-0.16
身高				0.50	-0.28
体重					-0.12

3. 分析方法及评价标准

本文的数据分析思路为：根据实际研究所获数据模拟出 100 个完整数据集，在此基础上，分别根据完整数据中第二次和第三次腰椎骨密度 T 值的数据缺失率（分别为 22.34% 和 28.37%）生成 100 个有数据缺失的数据集，然后再用各种缺失数据填补方法对缺失数据集进行填补，最后根据评价指标来评价各填补方法的优劣。

数据分析软件为 SAS 9.1，模拟数据集采用 IML 模块和 SAS 宏程序，缺失数据的处理和分析主要采用了 PROC MI 和 PROC MIANALYZE 过程。由于 REG 方法、PMM 方法和 PS 方法只能对单调缺失模式的数据进行填补，所以在用如上三种方法进行缺失数据填补时，本文首先对第二次腰椎骨密度 T 值填补 N（N = 3、5、10、15、20）次，使数据变成单调缺失后，再用相应的填补方法对第三次腰椎骨密度 T 值填补 1 次。

针对缺失数据填补效果优劣的评价指标本文采用准确度和稳定度[10]。对于变量 Y，100 个完整数据集有 100 个均数 $Y_1, Y_2, \cdots, Y_{100}$，这 100 个均数的平均值为 Y_{mean}，缺失数据经过处理后也会有 100 个均数 $\hat{Y}_1, \hat{Y}_2, \cdots, \hat{Y}_{100}$，均数的平均值为 $\hat{Y}_{mean}$，则准确度指标定义为：$BIAS_{mean} = \hat{Y}_{mean} - Y_{mean}$

$BIAS_{mean}$ 指标的绝对值越小说明估计均数时偏差越小，准确度越高。

稳定度指标定义为：$MSE_{mean} = \sum_{i=1}^{100} (\hat{Y}_i - \hat{Y}_{mean})^2$

MSE_{mean} 指标越大说明估计均数时稳定度越好[11]。

同理可以计算 100 个标准误的 $BIAS_{stderr}$ 和 MSE_{stderr}。

结　果

用不同填补方法对模拟的缺失数据集进行填补，第二次和第三次腰椎骨密度 T 值填补效果较好的前五位的评价指标结果分别如表 3 和表 4 所示：

表 3　不同填补方法对第二次腰椎骨密度 T 值填补效果

填补方式	E_{mean}	E_{stderr}	$BIAS_{mean}$	$BIAS_{stderr}$	MSE_{mean}	MSE_{stderr}
Ad Hoc	-1.5919	0.1691	0.0005	0.0440	2.4732	0.0118
LOCF	-1.6528	0.1250	-0.0604	-0.0001	1.6860	0.0027
PMM 3 次	-1.6046	0.1239	-0.0122	-0.0012	1.8019	0.0030
PMM 10 次	-1.5985	0.1242	-0.0061	-0.0009	1.7979	0.0029
PMM 15 次	-1.5936	0.1243	-0.0012	-0.0008	1.7990	0.0029
PMM 20 次	-1.5974	0.1244	-0.0050	-0.0007	1.7984	0.0029
PS 3 次	-1.5933	0.1428	-0.0009	0.0177	1.9023	0.0362
PS 5 次	-1.5958	0.1419	-0.0034	0.0168	1.8931	0.0280
PS 10 次	-1.5964	0.1401	-0.0040	0.0150	1.9466	0.0136
PS 15 次	-1.5946	0.1401	-0.0022	0.0150	1.9634	0.0112
PS 20 次	-1.5948	0.1401	-0.0024	0.0150	1.9107	0.0091
MCMC 3 次	-1.5936	0.1276	-0.0012	0.0025	1.8061	0.0036
MCMC 5 次	-1.5933	0.1273	-0.0009	0.0022	1.8170	0.0031
MCMC 10 次	-1.5930	0.1269	-0.0006	0.0018	1.8179	0.0029

*：REG 3 次表示用 REG 方法填补 3 次，其他类推。下划线表示填补效果最好的前五位。

从表 3 中可以看出对于第二次腰椎骨密度 T 值均数准确性的评价指标 $BIAS_{mean}$ 绝对值最小的前五位为：0.0005、0.0006、0.0009、0.0009、0.0012、0.0012 分别为 Ad Hoc 方法、MCMC 方法填补 10 次、MCMC 方法填补 5 次、PS 方法填补 3 次、MCMC 方法填补 3 次、PMM 方法填补 15 次。对于第二次腰椎骨密度 T 值均数稳定性的评价指标 MSE_{mean} 最大的前五位为：2.4732、1.9634、1.9466、1.9107、1.9023 分别为 Ad Hoc 方法、PS 方法填补 15 次、PS 方法填补 10 次、PS 方法填补 20 次、PS 方法填补 3 次。

对于第二次腰椎骨密度 T 值标准误准确性的评价指标 $BIAS_{stderr}$ 绝对值最小的前五位为：0.0001、0.0007、0.0008、0.0009、0.0012 分别为 LOCF 方法、PMM 方法填补 20 次、PMM 方法填补 15 次、PMM 方

法填补 10 次、PMM 方法填补 3 次。对于第二次腰椎骨密度 T 值标准误稳定性的评价指标 MSE_{stderr} 最大的前五位为：0.0362、0.0280、0.0136、0.0118、0.0112 分别为 PS 方法填补 3 次、PS 方法填补 5 次、PS 方法填补 10 次、Ad Hoc 方法、PS 方法填补 15 次。

表 4　不同填补方法对第三次腰椎骨密度 T 值填补效果

填补方式	E_{mean}	E_{stderr}	$BIAS_{mean}$	$BIAS_{stderr}$	MSE_{mean}	MSE_{stderr}
Ad Hoc	−1.8289	0.1652	−0.0044	0.0429	2.0351	0.0120
LOCF	−1.7780	0.1231	0.0465	0.0008	1.6278	0.0026
REG 3 次	−1.8265	0.1226	−0.0020	0.0003	1.6889	0.0028
REG 5 次	−1.8284	0.1227	−0.0039	0.0004	1.6976	0.0027
REG 10 次	−1.8364	0.1234	−0.0119	0.0011	1.6783	0.0028
REG 15 次	−1.8367	0.1233	−0.0122	0.0010	1.6825	0.0027
PS 3 次	−1.8259	0.1366	−0.0014	0.0143	1.7976	0.0266
PS 5 次	−1.8277	0.1393	−0.0032	0.0170	1.8190	0.0194
PS 10 次	−1.8314	0.1369	−0.0069	0.0146	1.8099	0.0110
PS 15 次	−1.8275	0.1356	−0.0030	0.0133	1.7753	0.0067

*：REG 3 次表示用 REG 方法填补 3 次，其他类推。下划线表示填补效果最好的前五位。

从表 4 中可以看出对于第三次腰椎骨密度 T 值均数准确性的评价指标 $BIAS_{mean}$ 绝对值最小的前五位为：0.0014、0.0020、0.0030、0.0032、0.0039 分别为 PS 方法填补 3 次、REG 方法填补 3 次、PS 方法填补 15 次、PS 方法填补 5 次、REG 方法填补 5 次。对于第三次腰椎骨密度 T 值均数稳定性的评价指标 MSE_{mean} 最大的前五位为：2.0351、1.8190、1.8099、1.7976、1.7753 分别为 Ad Hoc 方法、PS 方法填补 5 次、PS 方法填补 10 次、PS 方法填补 3 次、PS 方法填补 15 次。

从如上的结果综合来看，PS 方法填补 3 次在本文模拟的数据中填补效果最佳，而 MCMC 方法除在第二次腰椎骨密度 T 值的 $BIAS_{mean}$ 指标上表现较好外，在其他指标中都没有进入填补效果最好的前五位。

讨　论

在多重填补的四种方法里，PS 方法在第二次和第三次腰椎骨密度 T 值的 MSE_{mean} 指标和 MSE_{stderr} 指标上都有很好的效果，REG 方法在第三次腰椎骨密度 T 值的 $BIAS_{stderr}$ 指标上有很好的效果，PMM 方法在第二次腰椎骨密度 T 值的 $BIAS_{stderr}$ 指标上有很好的效果，而 MCMC 方法只在第二次腰椎骨密度 T 值的 $BIAS_{mean}$ 指标上有较好的效果。填补次数越多填补效果不一定越好。

因此从本文可以看出，对于任意缺失模式的缺失数据集，多重填补的 MCMC 并不是唯一的多重填补方法，采用单调缺失模式下的多重填补方法把任意缺失数据填补成单调缺失，在此基础上再进行一次该方法的填补，在某些条件下比 MCMC 填补的效果好。对于填补的次数并不是越多越好，而是要根据实际情况，进行数据模拟，从而找出最佳的填补次数。

A Simulated Comparison between Different Imputation Methods in Arbitrary Missing Data

Zhang Qiao, Li Ning, Zhang Qiuju, et al. Department of Health Statistics, Harbin Medical University (150086), Harbin

【Abstract】 **Objective** To evaluate the imputation effect of different imputation methods in arbitrary missing data. **Methods** First of all, we use the IML model in SAS software to simulate arbitrary missing data, which is about the calcium requirements and dietary evaluation of postmenopausal women in the north of china. Imputing the missing data through the MI and MIANALYZE processes. Accuracy and stability were used for the evaluation indices to compare the imputation effect of different methods. **Results** The effect of PS method when imputing 3 times is the best in this data, while the effect of MCMC method is not ideal. **Conclusion** The MCMC is not the unique multiple imputation method when imput arbitrary missing data. The PS, PMM, REG methods could turn the arbitrary missingness pattern into monotone missingness pattern, then we use the same method to imput once again. It is also an alternative imputation method.

【Key words】 Missing data; Arbitrary missingness pattern; Multiple imputation; Data simulation

参 考 文 献

1. Amold AM, Kronmal RA. Multiple imputation of baseline data in the cardiovascular health study. American Journal of Epidemiology, 2003, 157(1): 74-84.
2. Abraham, Todd W, Russell, et al. Missing data: a review of current methods and applications in epidemiological research. Current Opinion in Psychiatry, 2004, 17(4): 315-321.
3. James M, Robins, Wang N. Inference for imputation estimators. Biometrika, 2000, 87(1): 113-124.
4. Little RJ, Rubin DB. Statistical Analysis with Missing Data. New York: John Wiley&Sons, 1987.
5. Little RJ, Rubin DB. Statistical Analysis with Missing Data. 2nd ed. Hoboken, NJ: John Wiley&Sons, 2002.
6. 曹阳，谢万军，张罗漫. 多重填补的方法及其统计推断原理. 中国医院统计，2003，10(2)：77-81.
7. 李新华，夏结来. 多重填补处理有缺失数据的 2×2 交叉设计资料的应用. 2004 中国卫生统计学术会议论文集，2004：181-187.
8. 张熙，林燧恒. 多重填补在随机干预实验研究中的应用. 中国卫生统计，2011，28(5)：537-539.
9. 茅群霞. 缺失值处理统计方法的模拟比较研究及应用. 四川大学硕士毕业论文，2005.
10. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods, 2001, 6(4): 330-351.
11. 李宁. 钙干预试验骨密度缺失值的填补研究. 哈尔滨医科大学硕士毕业论文，2010.

（责任编辑：郭海强）