对模拟纵向数据集缺失值处理的几种方法比较

易昆南，袁中萸

（中南大学 数学科学与计算技术学院，湖南 长沙 410075）

摘 要：运用SAS9.0、数据模拟技术，分别模拟纵向完整数据集和具有各种缺失率的随机缺失数据集，采用多重填补法（MI）、期望值最大化法（EM）和回归插补法（Regression）对各缺失数据集进行处理，对结果进行比较和分析。结果表明，对不同缺失率的数据集，MI、EM和Regression法对缺失值的处理各有优劣。

关键词：多重填补法；期望值最大化法；回归插补法；缺失值

中图分类号：O212 **文献标识码**：A **文章编号**：1673-9833(2008)02-0048-04

Comparison on Several Methods in Simulated Longitudinal Data with Missing Values

Yi Kunnan，Yuan Zhongyu

（School of Mathematical Science and Computer Technology，Center South University，Changsha 410075，China）

Abstract: The simulated datasets with vary missing rates are treated by multiple imputation (MI)，expectation maximization（EM）and regression methods and the results are compared with that of complete dataset by running SAS 9.0 procedures. The results showed that in different missing rate data，MI，EM and Regression have their own advantages and disadvantages.

Key words: multiple imputation; expectation maximization; regression imputation; missing values

0 引言

在调查研究中，数据缺失是一个常见的问题。数据缺失可能导致样本信息减少、检验效能降低[1]以及增加统计分析的复杂性。当缺失数据过多时，可能完全失去利用价值，即使数据缺失在能够处理的范围之内，如果处理方式不恰当，可能造成分析结果的偏性或不能充分利用资料信息。

缺失值处理问题涉及的统计方法较多，不同方法对特定资料缺失值处理的优劣只有通过比较和鉴别才能显现。本研究拟采用数据模拟技术，比较多重填补法（multiple imputation，简称MI）、期望值最大化法（expectation maximization，简称EM）和回归插补法（regression imputation，简称Regression）3种缺失值处理方法的优劣。

1 数据模拟

通过SAS9.0编程，模拟一个完整数据集，该数据集中包含的观察数为n=100，1个因变量y，6个自变量，其中$x_1, x_2, x_3, x_4, x_5, x_6$均为连续变量，每一个体重复测量10次。对该数据集建立多元线性回归模型

$$y=\alpha+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\varepsilon,$$

估计该模型各参数及其标准误差。

对该完整数据集模拟100次，得到该模型各参数及其标准误差的平均值作为比较的标准。在此基础上，仍采用SAS9.0编程，对该完整数据集构造各种不同缺失率的随机缺失[2,3]数据集，对每一种缺失率的数据集均模拟100次，分别采用MI、EM与Regression法对每一种缺失率的数据集缺失值进行处理，得到上述模型各参数及其标准误差的估计值，并与完整相应参

收稿日期：2008-01-10

基金项目：湖南省自然科学基金资助项目（03JJY4071）

作者简介：易昆南（1954-），男，湖南长沙人，中南大学教授，硕士生导师，主要从事随机数学与建模方面的教学与研究.

数及其标准误差进行比较。

2　缺失值处理

多重填补法（MI法）由Rubin于1987年最早提出，它是一种用2个或更多的可得到的并能反映数据本身分布概率的值来填补缺失值的方法[4]。近年来，MI法在国外发展成为处理缺失值的最常用方法之一[5]，它在沿袭传统填补方法的基础上，结合数据收集者的专业背景来反映缺失数据的不确定性，从而使得填补结果更接近“真实”。但MI法也有其应用条件与适用范围，它要求数据呈随机缺失的形式，对数据的填补是MI法过程中的关键步骤。对每一个缺失的数据，MI法填补m（m>1）次，这样，第一次填补就产生第一个完全数据集，以此类推，将产生m个完全数据集[6]。对每一个完全数据集，都采用标准的完全数据分析方法进行分析，并将所得结果进行综合，得到最终的统计推断。

期望值最大化法（EM法）反复强调采用先估计缺失值，然后估计参数。“M”的步骤是假设没有缺失数据而进行最大似然估计，然后进行“E”步，即是在给定的观测数据和当前得到的参数估计值的条件下，求出缺失值的条件期望，缺失数据可以用期望值替代，继续以上步骤，直到参数的估计值收敛为止。

回归插补法（Regression法）是运用回归技术来替代缺失值的方法，它是通过多元回归方法建立变量关于数据集所有其它变量的回归模型，并用非标准化的结果预测该变量的缺失值来实现的。

3　结果

取模拟100次结果的各变量参数及其标准误差的均数，结果见表1~5。

表1　完整数据集各变量参数及标准误差

Table 1　Complete datasets of variable parameters and standard errors

取值项	变量参数符号						
	α	β_1	β_2	β_3	β_4	β_5	β_6
参数值	−31.843 68	0.611 14	−0.445 27	0.422 19	−0.047 37	−0.390 02	0.224 67
标准误差	7.649 37	0.073 27	0.159 97	0.062 36	0.053 66	0.189 36	0.124 48

表2　缺失率为10％的数据集经处理后各变量参数及标准误差

Table 2　The processed datasets with missing rate of 10% after each variable parameter and standard error

处理方法	变量参数及标准误差							
	α		β_1		β_2		β_3	
	参数值	标准误差	参数值	标准误差	参数值	标准误差	参数值	标准误差
EM法	−39.131 73	10.728 41	6.840 23	0.904 24	−0.371 82	0.212 58	0.297 81	0.078 70
Regression法	−39.187 89	2.995 86	6.011 01	0.289 94	−0.519 66	0.075 99	0.420 67	0.021 01
MI法填补3次	−26.65197	6.052 41	5.225 82	0.583 47	−0.377 12	0.173 73	0.755 11	0.056 42
MI法填补5次	−27.37331	4.549 27	5.178 55	0.487 82	−0.347 74	0.131 61	0.742 17	0.042 23
MI法填补10次	−26.44991	3.218 21	5.238 10	0.307 32	−0.369 20	0.090 89	0.755 99	0.030 06

处理方法	变量参数及标准误差					
	β_4		β_5		β_6	
	参数值	标准误差	参数值	标准误差	参数值	标准误差
EM法	−0.157 37	0.078 24	−0.223 90	0.265 44	0.174 25	0.149 88
Regression法	0.015 69	0.027 33	−0.194 47	0.095 64	0.334 18	0.050 67
MI法填补3次	−0.099 35	0.045 58	−0.777 10	0.174 57	0.338 98	0.127 69
MI法填补5次	−0.097 84	0.034 96	−0.784 71	0.133 82	0.329 36	0.096 46
MI法填补10次	−0.101 21	0.023 90	−0.802 48	0.092 19	0.336 53	0.067 98

表 3 缺失率为 20% 的数据集经处理后各变量参数及标准误差

Table 3 The processed datasets with missing rate of 20% after each variable parameter and standard error

处理方法	变量参数及标准误差							
	α		β_1		β_2		β_3	
	参数值	标准误差	参数值	标准误差	参数值	标准误差	参数值	标准误差
EM 法	−21.221 0	1.58743	4.746 05	0.308 85	−0.023 58	0.074 17	0.466 96	0.011 50
Regression 法	−17.298 8	19.530 08	2.226 69	1.192 34	0.317 00	0.332 08	0.531 20	0.129 98
MI 法填补 3 次	−30.623 5	12.761 70	0.520 11	0.770 76	−0.556 23	0.183 87	0.784 51	0.103 47
MI 法填补 5 次	−29.877 4	9.337 93	0.932 91	0.568 44	−0.391 20	0.133 20	0.330 48	0.075 64
MI 法填补 10 次	−20.622 4	6.043 09	3.191 69	0.360 53	−0.198 41	0.088 23	0.239 58	0.050 03

处理方法	变量参数及标准误差					
	β_4		β_5		β_6	
	参数值	标准误差	参数值	标准误差	参数值	标准误差
EM 法	−0.230 42	0.016 75	−0.896 95	0.065 79	0.460 95	0.035 78
Regression 法	−0.289 76	0.121 29	−0.775 74	0.488 79	0.620 82	0.389 43
MI 法填补 3 次	−0.078 45	0.093 56	−0.251 00	0.247 92	0.152 60	0.206 15
MI 法填补 5 次	−0.049 01	0.069 56	−0.542 36	0.182 03	0.210 01	0.150 00
MI 法填补 10 次	−0.057 96	0.043 81	−0.011 09	0.117 76	0.057 70	0.094 82

表 4 缺失率为 40% 的数据集经处理后各变量参数及标准误差

Table 4 The processed datasets with missing rate of 40 % after each variable parameter and standard error

处理方法	变量参数及标准误差							
	α		β_1		β_2		β_3	
	参数值	标准误差	参数值	标准误差	参数值	标准误差	参数值	标准误差
EM 法	12.630 2	6.891 01	2.567 61	0.893 84	−0.931 00	0.056 33	0.966 11	0.081 68
Regression 法	38.676 8	7.520 35	2.056 87	2.606 52	−0.684 85	1.221 07	0.397 19	0.244 31
MI 法填补 3 次	−39.612 5	13.927 17	1.581 30	0.771 38	−0.207 97	0.122 34	0.161 15	0.106 54
MI 法填补 5 次	−25.020 9	11.059 17	1.802 51	0.602 74	−0.133 23	0.097 61	0.084 45	0.080 83
MI 法填补 10 次	−32.628 8	7.665 89	1.029 70	0.418 22	−0.408 03	0.070 04	0.219 20	0.054 72

处理方法	变量参数及标准误差					
	β_4		β_5		β_6	
	参数值	标准误差	参数值	标准误差	参数值	标准误差
EM 法	0.017 34	0.009 72	0.361 81	0.205 99	0.016 37	0.011 55
Regression 法	−0.185 54	0.243 45	0.185 52	1.497 43	0.695 18	0.831 79
MI 法填补 3 次	−0.111 99	0.063 69	−0.146 58	0.240 30	0.027 12	0.120 59
MI 法填补 5 次	−0.119 33	0.053 08	0.195 13	0.186 93	0.026 96	0.093 68
MI 法填补 10 次	−0.091 31	0.035 22	−0.279 54	0.128 40	0.235 84	0.066 13

表5　缺失率为50％的数据集经处理后各变量参数及标准误差

Table 5　The processed datasets with missing rate of 50% after each variable parameter and standard error

处理方法	变量参数及标准误差							
	α		β_1		β_2		β_3	
	参数值	标准误差	参数值	标准误差	参数值	标准误差	参数值	标准误差
EM法	−33.751 70	14.371 82	4.685 78	0.788 50	−0.000 05	0.103 32	0.532 50	0.099 29
Regression法	−12.865 30	51.451 85	3.941 31	1.561 67	−0.052 24	0.369 63	0.074 97	0.198 32
MI法填补3次	−33.482 76	6.181 34	4.616 19	0.380 02	0.032 03	0.043 10	0.524 43	0.049 00
MI法填补5次	−34.168 40	4.890 78	4.872 90	0.306 24	0.010 90	0.033 99	0.511 24	0.038 93
MI法填补10次	−32.478 64	3.429 04	4.699 37	0.202 88	−0.022 87	0.023 11	0.533 57	0.025 15

处理方法	变量参数及标准误差					
	β_4		β_5		β_6	
	参数值	标准误差	参数值	标准误差	参数值	标准误差
EM法	−0.000 06	0.054 44	−0.897 38	0.172 01	0.001 27	0.240 67
Regression法	0.045 47	0.249 71	−0.002 83	0.578 31	−1.465 85	0.959 06
MI法填补3次	−0.013 04	0.023 58	−0.859 99	0.085 13	−0.155 08	0.052 20
MI法填补5次	−0.006 31	0.018 35	−0.883 76	0.066 58	−0.135 37	0.042 77
MI法填补10次	−0.004 01	0.012 50	−0.889 47	0.044 65	−0.136 99	0.028 21

当缺失率为10％时，用EM法与Regression法进行处理的结果差别不太，MI方法与之相比，略有优越性，处理后的结果与完整数据集的分析结果稍接近点。当数据缺失率在20％~40％时，MI法处理后的结果较EM法与Regression法更接近完整数据的分析结果，且当缺失率较低时，填补5次的结果比较接近完整数据分析结果；而缺失率较高时，填补10次的结果相对较优。当缺失率较低时，EM法稍优于Regression法；缺失率较高时，Regression法又表现出比EM法更优的一面。当缺失率达50％时，3种方法处理结果均不理想，相对而言，EM法与Regression法处理后的结果稍好一点，但其中个别的参数及其标准误差的估计与完整数据分析结果差别还是较大的。

4　结论

1）当数据呈随机缺失且缺失率≤10％时，MI法处理的结果与“实际情况”较接近，而EM法与Regression法处理结果差别不大且效果不太好。

2）当数据缺失率在20％~40％时，MI法处理缺失数据显示了其优越性，填补结果较其它两种方法更接近“实际情况”。另外，当缺失率相对较低时，MI处理数据只需较少的填补次数即可达到较好的效果；当缺失率相对较高时，需增加填补次数，以提高填补效率。

3）当数据缺失率相对较低时，EM法比Regression法效果好；当缺失率相对较高时，Regression法比EM法效果好。

4）当缺失率达50％时，不管哪种方法处理，结果均不理想，相比之下，EM法比Regression法稍优。这说明，当数据缺失率太高时，数据即失去了可利用价值，先进和复杂的填补方法也不能产生理想的结果。

参考文献：

[1] Fairclough Diane L，Peterson Harriet F，Chang Victor. Why are missing quality of life data a problem in clinical trials of cancer therapy[J]. Statistics in Medicine，1998，17：667-677.

[2] Little R J A，Rubin D B. Statistical Analysis with Missing Data[M]. New York：John Wiley&Sons，1987.

[3] Abraham，Todd W，Russell，et al. Missing data；a review of current methods and applications in epidemiological research [J]. Current Opinion in Psychiatry，2004，17(4)：315-321.

[4] Rubin D B. Multiple Imputation for Nonresponse in Surveys [M]. New York：John Wiley&Sons，1987.

[5] Robins James M，Wang Naisyin. Inference for imputation estimators[J]. Biometrika，2000，87(1)：113-124.

[6] Buuren S Van，Boshuizen HC，Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis[J]. American Journal of Epidemiology，2003，157(1)：78-84.

（责任编辑：张亦静）