·论 著·

多变量缺失数据的不同处理方法及分析结果比较

武建虎[1], 贺 佳[1*], 贺宪民[1], 程红岩[2]
(1. 第二军医大学卫生勤务学系卫生统计学教研室, 上海 200433; 2. 东方肝胆外科医院放射科, 上海 200438)

[摘要] 目的: 探讨多变量缺失数据的不同处理方法对结果的影响。方法: 分别利用删除含缺失值的观察、简单填补、多重填补3种方法对多变量中度缺失的925例肝癌患者的临床资料进行统计分析并对其结果进行比较。结果: 不同方法所产生的结果差别较大。在$\alpha=0.05$的水平下, 利用多重填补处理的数据集分析得到影响肝癌患者生存时间的危险因素: 临床分期、肝硬化史、门脉癌栓、g-GT和WBC; 而用删除含缺失值方法得到的却是: TNM分期、碘油剂量、AST、ALP; 简单填补比多重填补多产生3个危险因素, 分别是: TNM分期、ALP和AFP。结论: 本资料采用删除含缺失值的观察的方法结果最差; 简单填补相对较好, 但容易降低标准误、减小P值; 而多重填补处理比较合理、科学。建议对多变量数据缺失的处理一定要慎重。

[关键词] 多变量; 缺失值; 多重填补; 肝肿瘤
[中图分类号] R 311; R 735.7 [文献标识码] A [文章编号] 0258-879X(2004)09-1013-04

Comparison of different methods in management of multivariate missing data

WU Jian-Hu[1], HE Jia[1*], HE Xian-Min[1], CHENG Hong-Yan[2] (1. Department of Health Statistics, Faculty of Health Services, Second Military Medical University, Shanghai 200433, China; 2. Department of Radiology, Eastern Hepatobiliary Surgery Hospital, Shanghai 200438)

[ABSTRACT] **Objective**: To explore the results of different methods for managing multivariate missing data. **Methods**: Case deletion, simple imputation and multiple imputation were compared when used for analyzing the clinical data of 925 liver cancer patients with medium multivariate missing data. **Results**: There were differences among the 3 methods. When $\alpha=0.05$, the risk factors influencing patients' survival time were clinical staging, history of hepatic cirrhosis, portal vein tumor thrombas, and levels of g-GT and WBC with multiple imputation, and were TNM staging, lipiodol dose, AST and ALP with case deletion. The 3 more factors of simple imputation were TNM staging, ALP and AFP compared with multiple imputation. **Conclusion**: Simple imputation is superior to case deletion in management of multivariate missing data but tends to make standard error smaller and P value lower. Multiple imputation is more reasonable and scientific than the other 2 methods.

[KEY WORDS] multivariate; missing data; multiple imputation; liver neoplasms
[Acad J Sec Mil Med Univ, 2004, 25(9): 1013-1016]

在临床资料中, 由于种种原因存在有缺失值的现象。缺失数据带来的主要问题有: 数据中的信息不能被完全提取、数据的处理与分析复杂以及容易产生偏倚[1,2]。

对缺失数据的处理常用的方法是将含缺失数据的观察剔除, 或者采用简单填补即为每一个缺失值填补一个替代值。20多年前有学者提出了多重填补的方法[3,4], 并成为处理缺失数据强有力的工具。本研究采用该3种方法处理缺失数据, 并对其填补效果进行评价。

1 资料和方法

1.1 资料来源 本研究资料来自东方肝胆医院1996年12月至2002年12月经放射介入治疗(TAE)的肝癌患者, 共收集有确切死亡时间的患者为925例, 经单因素分析并结合实际筛选出了对患者生存时间有影响的变量。数据中多变量(27个)的变量值及缺失情况如表1。平均缺失为9.62%(0.76%～24.86%)。不含缺失值的完全观察有220例, 占全部例数的23.78%。

1.2 删除含有缺失值的观察(case deletion) 在分析时将含有缺失值的观察整个排除在分析之外。

1.3 简单填补(simple imputation) 给每一个缺失值填补一个替代值, 对于填补后的数据集可以使用针对完整数据集分析的方法进行分析。常用的方法有均数(中位数)替代、回归方法、Hot deck方法等。本研究选用了最常用的均数替代, 即对计量变量中分布为偏态的以中位数填补, 正态或近似正态以均数填补, 分类、等级和二值变量以出现频数最多的数值填补。

* [作者简介] 武建虎(1974-), 男(汉族), 硕士生
E-mail: wjh204041@sina.com
*Corresponding author. E-mail: hejia@smmu.edu.cn

表 1　925 例肝癌患者多因素分析入选变量及缺失情况列表

Tab 1　Missing status of selected variables on multivariate analysis of 925 liver cancer patients

Variable	Label	Type	Missing rate(%)
$x1$	Sex	Binary	1.3
$x2$	Work category	Qualitative	1.73
$x3$	Clinical stage	Ranked	3.78
$x4$	Ways of discovering disease	Qualitative	4.43
$x5$	Hepatitis history	Qualitative	2.7
$x6$	Clinical history of liver cirrhosis	Binary	4.86
$x7$	Family history for liver cancer	Binary	2.27
$x8$	Family history for malignant tumor	Binary	3.14
$x9$	Child-Pugh scores	Ranked	7.24
$x10$	Tumor number	Binary	19.89
$x11$	Portal tumor thrombus in B-ultrsonography	Ranked	7.68
$x12$	TNM stage	Ranked	17.3
$x13$	Lipiodol dose	Quantitative	2.7
$x14$	CBP	Quantitative	26.27
$x15$	ALB	Quantitative	0.76
$x16$	A/G	Quantitative	12.22
$x17$	ALT	Quantitative	2.7
$x18$	AST	Quantitative	18.38
$x19$	G-GT	Quantitative	18.92
$x20$	WBC	Quantitative	9.84
$x21$	Pt	Quantitative	6.16
$x22$	AFU	Quantitative	24.86
$x23$	ALP	Quantitative	6.38
$x24$	AFP	Quantitative	7.14
$x25$	CEA	Quantitative	22.27
$x26$	Age	Quantitative	6.7
y	Survival time	Quantitative	0

1.4　多重填补(multiple imputation, MI)　多重填补为每一个缺失值产生一套可能的填补值, 每一个值用来填补数据集中缺失, 产生有代表性的 5～10 套完整数据集。然后对每一个填补数据集都用针对完整数据集的统计方法进行统计分析。最后将各个数据集分析结果进行综合, 产生最终的统计推断。主要包括数据填补和综合推断两部分。

1.4.1　数据填补(data imputation)的过程　使用马尔科夫链蒙特卡罗(Markov chain Monte Carlo, MCMC)法, 填补前对明显偏态的变量进行最大正态化, 使数据服从多元正态分布; 对分类变量 $x2$、$x4$、$x5$ 采用出现频率最高的水平填补后以哑变量形式引入填补模型[5]。主要过程为利用期望最大(expectation maximization, EM)法则获得参数估计, 在一个假定的模型参数下对缺失值获得一个预测值, 然后基于极大似然估计利用原有数据和预测值得到一个新的模型参数, 如此反复迭代直至收敛为止。根据 EM 法获得的模型参数对缺失值随机填补, 然后, 从基于原有数据和填补值的贝叶斯后验分布中抽取新的参数。这样交替地模拟参数和填补值会产生一个足够长的马尔科夫链并最终稳定, 此时, 参数的分布稳定为后验分布, 而缺失数据的分布稳定为一个可产生预测的分布, 然后就可以近似独立地从该分布中为缺失值抽取填补值。当所有的缺失值都被替代后, 就形成一个完整的数据集, 重复此过程 m 次形成 m 个完整数据集。本研究重复 3 次形成 3 个独立的数据集 a、b 和 c。

1.4.2　综合推断　填补后的数据集经 SAS 分析后将参数估计、标准误及协方差矩阵输入 NORM 软件。依据 Rubin 提出的法则, 假定 $\hat{Q}_j$ 和 U_j 为第 j 个数据集分析得到的总体参数的点估计值和方差估计值, 整体参数估计值就是单个估计值的均数:

$$\bar{Q} = \frac{1}{m}\sum_{j=1}^{m}\hat{Q}_j$$

考虑到填补后数据的变异来自填补数据集间的变异和填补数据集内的变异, 因此总体方差的估计由 2 部分组成, 填补内方差($\bar{U}$)和填补间方差(B):

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m}U_j$$

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{Q}_j - \bar{Q}\right)^2$$

总体方差的估计 T 是 $\bar{U}$ 和 B 的校正值之和:

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B$$

$\sqrt{T}$ 就是参数的总的标准误。可以看出当没有缺失数据时, $\hat{Q}_1$、$\hat{Q}_2$、…$\hat{Q}_m$ 都是一样的, $B = 0$, $T = \bar{U}$。因此, 从方差的角度看, B 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。$\sqrt{T}$ 就是参数的总的标准误, 对参数的显著性检验($H_0: Q = 0$)通过计算, $t = \bar{Q}/\sqrt{T}$, t 服从近似 t 分布, 自由度为:

$$df = (m-1)\left[1 + \frac{m\bar{U}}{(m+1)B}\right]^2$$

对于多因素综合推断($H_0: Q_1 = Q_2 = \cdots = Q_k = 0$, $Q_1, Q_2, \cdots, Q_k$ 为参数), $\hat{Q}_j$ 和 U_j 为第 j 个数据集分析得到的总体参数的点估计值向量和方差矩阵, 计算检验统计量 F。上述过程由 Schafer JL 编制的基于 Windows95/98/NT 的 NORM 2.03 软件实现。

1.5　统计学处理　填补后的数据集采用 COX 比例风险模型进行分析, 主要目的是研究 TAE 患者生存时间的影响因素。对分类变量设置为哑变量进入 COX 模型, 使用 SAS 8.2 软件的 Phreg 过程完成。

2　结　果

对 3 种方法的处理结果整理分别见表 2～4, 其中 3 个分类变量 $x2$、$x4$、$x5$ 分别设置为哑变量 $x2$-1～$x2$-5、$x4$-1～$x4$-3、$x5$-1～$x5$-3 进入模型, X-1 (X 代表 $x15$、$x16$、…、$x25$ 等变量) 表示变量进行正态转化后的变量, 参数估计正值表示危险因素, 负值为保护因素。对于前 2 种方法均采用逐步筛选法, 入选和剔除标准均设为 0.05。为便于对照比较将多重填补综合推断的 $P < 0.15$ 的结果均输出。3 种方法处理模型均有意义 ($P < 0.001$), 在 $\alpha = 0.05$ 的水平下, 3 种方法筛选的变量有差别。本研究资料中部分变量缺失较多, 对这类数据采用删除含缺失的观察处理是不合适的, 与多重填补相比有明显的不同。简单填补得到的结果与多重填补相比标准误减低、P 值减小, 具有统计学意义的变量增多 (如变量 $x23$、$x24$ 和 $x12$), 但与删除含缺失值的观察相比已没有变量 $x17$ 和 $x18$, 说明本资料使用简单填补要略好于第 1 种方法。多重填补的结果是 $x3$、$x6$、$x19$、$x20$、$x11$ 为危险因素, 而 $x7$ 为保护因素。本研究还对多重填补的 3 个数据集分别进行 COX 分析, 分析结果见表 5, 从中可看出 3 个数据集分析的结果有差异, 参考表 4 的缺失信息率, 该指标是根据数据缺失造成的方差增量来计算的, 反映了由于缺失数据引起的统计上的不确定性。

表 2　220 例完全观察 COX 分析结果

Tab 2　COX analysis of 220 complete cases

Variable	Parameter estimate	s	χ^2	P	Hazard ratio
$x17$-1	-0.329 98	0.132 81	6.172 9	0.013	0.719
$x18$-1	0.396 59	0.132 61	8.943 7	0.002 8	1.487
$x23$-1	0.054 73	0.015 84	11.935 4	0.000 6	1.056
$x13$	0.190 53	0.069 61	7.492	0.006 2	1.21
$x12$	0.180 09	0.062 16	8.393 5	0.003 8	1.197

表 3　简单填补后 COX 分析结果

Tab 3　COX analysis with simple imputation

Variable	Parameter estimate	s	χ^2	P	Hazard ratio
$x3$	0.331 5	0.114 14	8.435 3	0.003 7	1.393
$x6$	0.220 09	0.087 1	6.384 5	0.011 5	1.246
$x7$	-0.319 95	0.131 82	5.890 7	0.015 2	0.726
$x19$-1	0.025 15	0.007 29	11.908 4	0.000 6	1.025
$x20$-1	0.303 91	0.079 45	14.633 1	0.000 1	1.355
$x23$-1	0.019 73	0.006 55	9.068 7	0.002 6	1.02
$x24$-1	0.042 97	0.015 45	7.741 7	0.005 4	1.044
$x11$	0.197 27	0.034 84	32.059 6	0.000 1	1.218
$x12$	0.084 21	0.031 32	7.228	0.007 2	1.088

表 4　多重填补分析结果列表

Tab 4　Analysis with multiple imputation

Variable	Parameter estimate	s	t	P	Fraction missing information
$x3$	0.319 001	0.118 243	2.7	0.007 1	0.054 136
$x5$-1	-0.193 4	0.105 495	-1.83	0.070 2	0.169 619
$x6$	0.238 829	0.109 833	2.17	0.036 7	0.284 033
$x7$	-0.367 59	0.146 315	-2.51	0.013	0.121 517
$x19$-1	0.020 217	0.008 217	2.46	0.014 3	0.076 295
$x20$-1	0.286 233	0.084 774	3.38	0.001 2	0.188 045
$x23$-1	0.014 266	0.007 402	1.93	0.054	0.019 286
$x24$-1	0.034 982	0.019 418	1.8	0.084 8	0.350 158
$x11$	0.210 691	0.036 202	5.82	< 0.000 1	0.059 599
$x12$	0.062 776	0.034 755	1.81	0.072 2	0.100 871
$x13$-1	0.261 646	0.163 719	1.6	0.110 1	0.025 458

表 5　多重填补数据集 a、b 和 c 的 COX 分析结果

Tab 5　COX analysis with multiple imputation data set a, b, c

Variable	P (a)	P (b)	P (c)
$x3$	0.000 8	0.003 2	0.003 6
$x5$-1	0.030 6	> 0.15	0.052
$x6$	0.005 1	> 0.15	0.021 8
$x7$	0.023 7	0.016 2	0.005 7
$x9$	0.050 6	0.025 9	0.070 8
$x19$-1	0.012 9	0.003 4	0.005 3
$x20$-1	0.000 1	0.000 1	< 0.000 1
$x23$-1	0.068 1	0.011 8	0.036 9
$x24$-1	0.002 9	0.030 9	0.041 7
$x11$	< 0.000 1	< 0.000 1	< 0.000 1
$x12$	0.012 1	0.013 8	0.062 4
$x13$-1	0.070 9	0.028	0.071
$x14$	> 0.15	0.126 2	0.100 9
$x2$-4	> 0.15	0.135 4	> 0.15

3　讨　论

在临床资料的统计分析中, 很多人没有意识到缺失值所带来的影响。删除含有缺失值的观察在实践中应用最广, 如果样本量很大, 而缺失的数据量又很少时, 使用这种方法可以简化统计分析的过程, 但如果含有缺失值的观察很多, 使用该方法则不仅大大地减少了参与分析的样本量, 而且还会得出错误的推断。Scheffe[1] 认为数据缺失在 5% 以下并且完全观察的样本比例不低于 70%, 或者完全观察的样本是研究总体的 1 个随机样本时可以考虑使用此方法。

简单填补的优点是简单、容易操作, 适合于缺失量很小的数据, 缺点是导致标准误的降低和 P 值的减小, 使 I 型错误率升高, 还有容易引起系统偏倚,

而且该方法忽略了缺失数据预测的不确定性。

相反, 由 Rubin 提出的多重填补方法目前已成为处理缺失数据强有力的工具。本研究采用的 MCMC 法是数据模拟有效的手段, 利用它来进行填补是比较准确的, 而综合推断又结合了缺失数据预测的不确定性, 所以多重填补产生的关于标准误及 P 值的推论是有效的。

本研究表明 3 种方法处理结果有明显差异: 删除缺失值观察是最不可取的; 简单填补相对较好, 但易引起 P 值减小而且未考虑到缺失值预测的不确定性; 多重估算处理是有效的、合理的, 尤其数据缺失率继续增大时, 它的优势将更加突出[6]。本研究提示对于多变量缺失数据的处理在方法的选择上一定要谨慎。

[参 考 文 献]

[1] Scheffe J. Dealing with missing data[J]. *Res Lett Inf Math Sci*, 2002, 3: 153-156.

[2] 贺　佳, 陆　健 主编. 医学统计学中的 SAS 统计分析[M]. 上海: 第二军医大学出版社, 2002. 165-181.

[3] Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: a data analyst's perspective[J]. *Multivariate Behavioral Research*, 1998, 33(4): 545-571.

[4] 曹　阳, Ritu S, Ajay T. 居民健康调查资料中的缺失数据的多重估算[J]. 中国卫生统计, 2002, 19(5): 280-282.
Cao Y, Ritu S, Ajay T. Multiple imputation for missing data on household health survey data[J]. *Zhongguo Weisheng Tongji (Chin J Health Statistics)*, 2002, 19(5): 280-282.

[5] Darmawan I GN. NORM software review: handling missing values with multiple imputation methods[J]. *Evaluat J Australasia*, 2002, 2(1): 51-57.

[6] Bernards CA, Farmer MM, Qi K, *et al*. Comparison of two multiple imputation procedures in a cancer screening survey[J]. *J Data Sci*, 2003, 1(1): 1-20.

[收稿日期] 2004-01-04　　[修回日期] 2004-05-19
[本文编辑] 尹　茶

·个案报告·

下颌侧切牙双根管一例报告

Double root canal of mandibular lateral incisor: a case report

王　芳, 王忠亮 (山东省济宁口腔医院牙体牙髓病科, 济宁 272045)

[关键词] 下颌侧切牙; 双根管; 牙髓炎
[中图分类号] R 781.31　　[文献标识码] B　　[文章编号] 0258-879X(2004)09-1016-01

1　临床资料　患者, 男, 33 岁, 因下前牙疼痛 1 年于 2003 年 8 月 11 日来我院就诊。专科检查: 41、42 无龋, 切端磨损, 近髓, 探敏感, 叩不适, 温度测验和电测验反应迟钝。X 线片示 41、42 根周膜增宽, 42 根尖 1/2 根管影像不清晰(图 1A)。临床诊断: 41、42 慢性牙髓炎。治疗: 41、42 局麻下开髓、拔髓, 42 髓腔唇舌径较大, 髓底探及唇舌向两根管口。唇侧根管偏唇侧, 电测根管长度: 17 mm, 根管预备初尖锉为 10 号, 舌侧根管较偏舌侧, 与牙体长轴约呈 25 角, 电测根管长度: 16 mm, 根管预备初尖锉为 8 号。双根管采用逐步后退法根管预备, 根管内封 FC 棉捻。1 周后复诊, 疼痛消失, 无其他不适, 叩诊(-)。氧化锌丁香油糊剂+ 牙胶尖充填根管, 改变投射角度拍 X 线片示两根管均恰填(图 1B)。

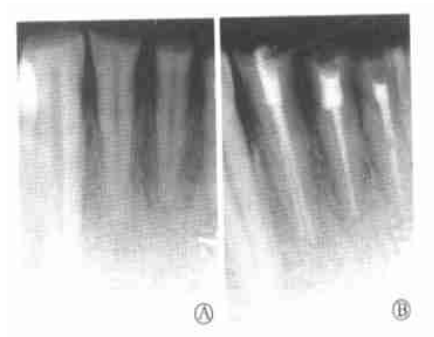

图 1　根充术前后 X 线片图像
A: 术前 42 根尖 1/2 根管影像不清晰;
B: 术后 42 双根管均恰填

2　讨　论　全口牙中, 下颌切牙髓腔体积最小, 根管多为扁而窄的单根管, 分为唇舌向两根管者较少, 约占 10%。双根管的下前牙在临床上也较为少见, 据报道下颌侧切牙双根管发生率为 9.2%～12.8%。在临床工作中, 双根管下前牙舌侧根管常因疏忽或不熟悉根管解剖而被遗漏, 因此在进行根管治疗时, 尤其对一些髓腔唇舌径较大的下颌侧切牙, 一定要有意识地寻找双根管, 以免遗漏, 造成根管治疗失败。牙髓病治疗后出现残髓炎或根尖周炎, 根尖周病治疗后不愈者, 应考虑是否有双根管遗漏根管。本例临床诊断为 41、42 慢性牙髓炎, 在治疗过程中发现 42 髓腔唇舌径较大, 髓底探及唇舌向两根管口, 遂对双根管进行治疗, 效果良好。

[收稿日期] 2004-01-08　　[修回日期] 2004-04-23
[本文编辑] 孙　岩

[作者简介] 王　芳(1978-), 女(汉族), 住院医师