多重填补的方法及其统计推断原理

曹 阳[1] 谢万军[2] 张罗漫[1]

【摘要】 **目的** 描述数据缺失的特征和数据缺失模式，对 Rubin 最早提出的多重填补(multiple imputation，MI)的基本概念、填补和分析缺失数据的方法、综合统计推断进行了探讨，分析了 MI 的特点、局限性以及应用 MI 方法处理不完整数据集时需要注意的地方。**方法** 通过计算机模拟，用 MI 方法将每一个缺失值用一系列可能的值填补，然后使用常规的、针对完全数据集的统计方法对多重填补后得到的若干数据集进行分析，并把所得的结果进行综合。**结果** 多重填补值显示出了缺失数据的不确定性，使得已有数据得到了充分利用，从而对总体参数做出了更为准确的估计。**结论** MI 方法为处理存在缺失值的数据集提供了有用的策略，并且适用于多种数据缺失的场合。

【关键词】 多重填补 缺失数据 广义线性模型 logistic 回归 马尔科夫链蒙特卡罗

中图分类号：R195.1 **文献标识码**：A **文章编号**：1006-5253(2002)04-0077-05

Methods of Multiple Imputation and Related Inference Theory *Cao Yang, Zhang Luoman. Department of Health Statistics, The Second Military Medical University, Shanghai, 200433. Xie Wanjun, Classification, Assessment and Survey Unit, WHO, Switzerland, Geneva.*

【**Abstract**】 **Objective** To describe the basic concepts, methods for imputing and analyzing missing data, combining inference of multiple imputation (MI) technique which was originally proposed by Rubin, and discuss limitations and cautions to consider when using MI as an analytic strategy for incomplete data settings. **Methods** By computer simulation, MI replaces each missing value with a set a plausible values. The multiply imputed data sets are then analyzed by using standard procedures for complete data and combining the results from these analyses. **Results** The multiply imputed data generated by MI procedure represent the uncertainty about the missing value and help us use data available to estimate population parameter more effectively. **Conclusion** MI provides a useful strategy for dealing with data sets with missing values and is broadly applicable to a variety of missing data settings.

【**Key words**】 Multiple imputation Missing data Generalized linear models Logistic regression Markov Chain Monte Carlo

数据缺失是在很多科学研究中经常出现的现象，对于统计分析人员来说，如何选择合适的方法对不完整的数据集进行分析，是一个棘手的问题。缺失数据带来的主要问题有：效率降低、数据的处理和分析复杂、观察到的数据与未观察到的数据间的差异所产生的偏倚。在近几十年，缺失数据的统计方法的发展是统计学研究中的一个活跃领域，20 多年前由 Rubin 提出来的多重填补(multiple imputation，MI) 方法就是其中一种，多重填补的步骤及其统计推断原理如图 1 所示：*

首先，MI 为每个缺失值产生一套可能的填补值，这些值反映了无响应值的不确定性，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，产生最终的统计推断，这一推断能够体现出由于数据填补而产生的不确定性。

MI 方法最初主要用于抽样调查和普查的大型数据集中。随着新的计算方法和统计软件的出现，该方法已被越来越多地应用于生物医学、行为学和社会科学领域[1~4]。

1 数据缺失的特征

在进行多重数据填补时，我们必须考虑数据是以怎样的形式缺失的，从而才能决定采用什么样的填补方法，下面对数据缺失的特征作一简单介绍[5]。

* 作者单位：1 200433 第二军医大学卫生统计学教研室 上海市
2 世界卫生组织分类、评估和调查研究组 瑞士日内瓦

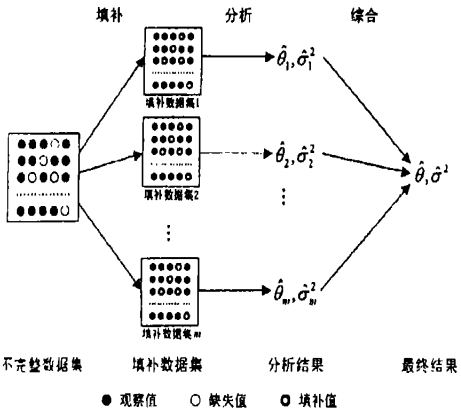

图 1　多重填补步骤及其统计推断原理

1.1　数据缺失机制(missingness mechanism)

现将数据集中不含缺失值的变量称为完全变量,这部分变量用 X_{obs}表示;数据集中含有缺失值的变量称为不完全变量,这部分变量用 X_{mis}表示。用向量 $R=(R_1, R_2, \cdots, R_n)$作为响应指示变量,当 $R_i=1$ 时,表示变量 X_i 被观察到(或被测量到),R_i 为 0 时,表示变量 X_i 为缺失值,则当满足:

$$P(R|X)=P(R|X_{obs}, X_{mis})=P(R|\phi) \quad (1)$$

其中 ϕ 是与数据集中任何变量都无关的参数。这种数据缺失机制被称为完全随机缺失(missing completely at random, MCAR)。

当 $P(R|X)=P(R|X_{obs}, \phi)$,即数据的缺失仅仅依赖于完全变量,这种数据缺失机制被称为随机缺失(missing at random, MAR)。在 MAR 情况下,不完全变量中的缺失值有可能系统性地大于或小于观察到的值。在 MI 方法中,都是假设数据是随机缺失的。

如果不完全变量中数据的缺失既依赖于完全变量,又依赖于不完全变量本身,这种缺失被称为不可忽略的(nonignorable)缺失。

1.2　数据缺失模式(missingness pattern)

假设数据集是由 n 个变量、p 个观测组成的 $p\times n$ 矩阵,对这个矩阵进行适当的行、列对换后,可以得到这样一个矩阵,它呈现出一种层级缺失的模式,即:当矩阵中的元素 X_{ij}缺失时,则对任意的 $k\geqslant i$ 和 $l\geqslant j$,元素 X_{kl}也是缺失的(见图 2)。这种数据缺失模式被称为单调缺失模式(monotone missingness pattern)。对于单调缺失模式来说,可以用一些简单的填补方法,不过在大多数复杂的调查中,这种缺失模式很少见。

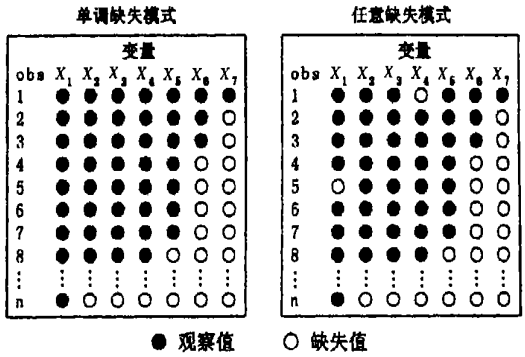

图 2　数据缺失模式

不满足单调缺失模式的,被称为任意缺失模式(arbitrary missingness pattern)。对这种缺失模式进行数据填补,方法较为复杂。如果可能的话,可以先把其中非单调缺失的观测值填上,把数据集换为单调缺失,从而就可以采用针对单调缺失模式的填补方法。

2　多重填补的方法

数据填补是多重填补统计分析中的关键一步,填补时一方面要考虑到填补的不确定性,同时还要考虑到观察到的完全变量与不完全变量间的关系。根据 MAR 假设,在以 X_{obs}为条件的基础上,X_{mis}的缺失是随机的,我们就可以从条件分布 $f(X_{mis}|X_{obs})$中产生填补值 $X^{(1)}, X^{(2)}, \cdots, X^{(m)}$。

对于单调缺失模式,有多种方法可以选用,如针对连续型变量的预测均数匹配(predictive mean matching, PMM)法、趋势得分(propensity score, PS)法;针对离散型变量的判别分析和 logistic 回归;对于复杂的缺失模式,可采用马尔科夫链蒙特卡罗(Markov Chain Monte Carlo, MCMC)方法[6,7]。

2.1　PMM 方法

PMM 方法假设在不完全变量与完全变量间存在着线性回归关系。例如变量 X_i 是一个存在着缺失值的不完全变量,用完全变量 X_1、X_2、…、X_{i-1}拟合模型:

$$E[X_i|\beta]=\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_{i-1}X_{i-1} \quad (2)$$

得到模型回归系数的参数的估计 $\beta=(\beta_0, \beta_1, \beta_2, \cdots, \beta_{i-1})$。在每一次填补中,从 β 的后验分布中随机抽取新的参数 β^*,计算:

$$X_i^*=\beta_0^*+\beta_1^* X_1+\beta_2^* X_2+\cdots+\beta_{i-1}^* X_{i-1}+\sigma^*\varepsilon \quad (3)$$

公式(3)中的 σ^* 为来自于模型的方差估计,ε 为模拟的正态随机误差。将缺失的 X_i 用数据集中

最接近于 X_i^* 的值填补，就是 PMM 方法。这种方法可以保证在正态性假设不成立的情况下，填补进较为适当的值。如果直接用 X_i^* 作为填补值，就称为回归方法(regression method)。

2.2 PS 方法

趋势得分中的“趋势”是指基于 X_{obs}的 X_{mis}中的数据缺失的条件概率，运用 PS 方法对单调缺失数据集进行填补的步骤如下：

(1)拟合 logistic 回归模型

$$\text{logit}(p_i)=\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_{i-1}X_{i-1} \quad (4)$$

公式(4)中 $p_i=Pr(R_i=0\mid X_1, X_2, \cdots, X_{i-1})$，$\text{logit}(p)=\log(p/(1-p))$。

(2)根据模型计算变量 X_i 上每个数据缺失的倾向性得分 $\text{logit}(\hat{p}_i)$，并根据该得分将所有的观测分组(可根据百分位数分为 4 组或 5 组，如果观测数量较多，可分为更多的组)。

(3)用近似贝叶斯 Bootstrap (approximate Bayesian Bootstrap，ABB)方法对各组中的缺失值进行填补。用 $X_i^{(n_0)}$ 表示某一组中 n_0 个 X_i 的观察值，$X_i^{(n_1)}$ 表示该组中 n_1 个 X_i 的缺失值，用有放回的抽样方法从 $X_i^{(n_0)}$ 中随机抽取 n_1 个数值对 $X_i^{(n_1)}$ 中的缺失值进行填充。

(4)重复这一过程，直至 X_{mis}中的每个变量都得到了填补。

2.3 MCMC 方法

MCMC 是贝叶斯推断中一种探索后验分布的方法，运用该方法对该缺失数据集填补可分为两步：

(1)填补步(Imputation step)

根据给定的均数向量 μ 和协方差矩阵Σ，从条件分布 $P(X_{mis}\mid X_{obs}, \phi)$中为缺失值抽取填补值。

假设 $\mu=[\mu_1', \mu_2']'$是两部分变量的均数向量，μ_1 是 X_{obs}的均数向量，μ_2 是 X_{mis}的均数向量。同时设定

$$\Sigma=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12}' & \Sigma_{22}\end{bmatrix} \quad (5)$$

其中Σ_{11}是 X_{obs}的协方差矩阵，Σ_{22}是 X_{mis}的协方差矩阵，Σ_{12}是 X_{obs}与X_{mis}间的协方差矩阵。在多元正态分布的假设下，当给定 $X_{obs}=x_1$ 时，X_{mis}的均数为：

$$\mu_{21}=\mu_2+\Sigma_{12}'\Sigma_{11}^{-1}(x_1-\mu_1) \quad (6)$$

其对应的条件协方差矩阵为

$$\Sigma_{221}=\Sigma_{22}-\Sigma_{12}'\Sigma_{11}^{-1}\Sigma_{12} \quad (7)$$

(2)后验步(Posterior step)

在每一次循环运算中，用上一次填补步中得到的 μ 和Σ作为后验总体的均数向量、协方差矩阵对参数 ϕ 进行模拟。

循环进行这两步过程，产生一个足够长的马尔科夫链：

$$(X_{mis}^{(1)}, \phi^{(1)}), (X_{mis}^{(2)}, \phi^{(2)}), K \quad (8)$$

当该链会聚在一个稳定的分布 $P(X_{mis}, \phi\mid X_{obs})$时，就可以近似独立地从该分布中为缺失值抽取填补值。

每次循环过程也可表述为：填补步用第 t 次循环得到的参数 $\phi^{(t)}$从分布 $P(X_{mis}\mid X_{obs}, \phi^{(t)})$中抽取 $X_{mis}^{(t+1)}$，后验步从分布 $P(\phi\mid X_{obs}, X_{mis}^{(t+1)})$中抽取 $\phi^{(t+1)}$。

3 多重填补的综合统计推断

无论采用哪种填补方法，都是将数据集填补 m 次($m>1$)，产生 m 个完整数据集，然后用任何针对完整数据集的方法来对它们进行分析[8,9]。

对于我们感兴趣的总体参数 θ 和 σ^2 来说，θ 和 $\hat{\sigma}^2$ 分别是它们的点估计值，则对每个填补数据集进行相同的分析，我们会得到 m 个 θ^1、θ^2、…、θ^m 和方差 $\hat{\sigma}_1^2$、$\hat{\sigma}_2^2$、…、$\hat{\sigma}_m^2$。MI 的参数估计是对上面的结果进行综合，对 θ 的估计为：

$$\theta=\frac{1}{m}\sum_{i=1}^{m}\theta_i \quad (9)$$

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，一个是填补数据集内的变异，故方差的估计是由两部分组成，一是填补内方差：

$$\sigma_W^2=\frac{1}{m}\sum_{i=1}^{m}\hat{\sigma}_i^2 \quad (10)$$

一是填补间方差：

$$\sigma_B^2=\frac{1}{m-1}\sum_{i=1}^{m}(\theta_i-\theta)^2 \quad (11)$$

方差的估计 σ_T^2 是 σ_W^2 与 σ_B^2 的校正值之和：

$$\sigma_T^2=\sigma_W^2+\left(1+\frac{1}{m}\right)\sigma_B^2 \quad (12)$$

σ_T^2 的平方根就是 θ 的总的标准误。我们可以看出，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：

$$\theta\pm t_\nu\sqrt{\sigma_T^2} \quad (13)$$

公式(13)中的自由度 ν 的计算方法为：

$$\nu=(m-1)\left(1+\frac{m\sigma_W^2}{(m+1)\sigma_B^2}\right)^2 \qquad (14)$$

对总体参数 θ 缺失的部分信息的估计是：

$$\gamma=\frac{r+2/(\nu+3)}{r+1} \qquad (15)$$

公式(15)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：

$$r=\frac{(1+m^{-1})\sigma_B^2}{\sigma_W^2} \qquad (16)$$

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。我们应注意到，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($\nu=\infty$)。自由度 ν 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 ν 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 $\left(1+\frac{\gamma}{m}\right)^{-1}$，进行 m 次填补对缺失信息弥补的相对效率见表 1。

表 1　多重填补的相对效率

m	γ				
	10%	20%	30%	50%	70%
3	0.967 7	0.937 5	0.909 1	0.857 1	0.810 8
5	0.980 4	0.961 5	0.943 4	0.909 1	0.877 2
10	0.990 1	0.980 4	0.970 9	0.952 4	0.934 6
20	0.995 0	0.990 1	0.985 2	0.975 6	0.966 2

4　讨论

在 MI 出现以前，列表式删除(list-wise deletion，即只删除分析变量内的缺失值)和单重填补(single imputation)是处理缺失值问题的主要方法。但是它们没有考虑到缺失数据的不确定性以及缺失数据与观察到的数据间可能存在的系统性差异，所以难以提供关于总体参数的准确估计。而 MI 很好地弥补了它们的缺陷，一些人甚至把 MI 看成是统计学的魔术，好象可以把信息从无变有，而正确的观点应该是把 MI 当作一种表现缺失数据不确定性的工具[10,11]。

MI 与期望最大化(expectation maximization，EM)法则和其他基于观察到的数据计算极大似然估计的方法相近似，只不过 MI 是通过蒙特卡罗方法而不是数学公式完成相应的工作。在大样本的情况下，次数足够多的 MI 得出的推断与直接用极大似然法所得出的几乎一样。

在应用 MI 时，都假设缺失机制是随机缺失，这种假设可以使我们很方便地避开一些复杂的概率模型，但是在很多场合，这种假设是不成立的，在不可忽略的缺失机制下，应用 MI 方法产生的估计会有明显的偏倚，这是统计分析人员需要注意的一点。

另外，在大多数 MI 方法中，我们必须根据某种概率分布的假设产生相应的填补数据，如 PMM 和 MCMC 方法都是以多元正态性假设为基础，这一方面使得 MI 方法的应用受到了局限，另一方面使填补过程趋于复杂和难于处理。有经验的统计分析员都知道，真正的数据很少服从那些方便的模型，在 MI 的大多数应用中，即使是最好的填补模型也只是真实情况的一种近似。但幸运的是，一些经验表明，MI 的推断结果对填补模型的依赖性并不很强，在少许偏离分布假设的情况下，统计推断也趋于稳健。例如对二分类变量或是等级分类变量进行填补时，把基于正态假设产生的连续型填补值按照最靠近的分类进行取整，常常也是可以接受的。当变量呈严重偏态分布时，可以对其进行变换，使其接近正态分布，然后再对产生的填补值进行反变换。

我们应该认识到，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。在许多场合，通过参数模拟或是多重填补进行统计分析都是可行的，比如通过加权的方法也能产生很好的参数估计。在全参数模型中，常常可以通过一些特殊的数学方法直接从不完全数据中计算出参数的极大似然估计，因为无需反复模拟，这些方法在某种程度上比 MI 更有效率[12,13]。实际上，只要有充分的时间和资源，对于任何特定的问题，人们可能都会找到比 MI 更好的统计方法。而在实际应用中，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，能够大大提高统计分析人员的工作效率。

参　考　文　献

1　Barnard J，Meng X L. Applications of multiple imputation in medical study：from ADIS to NHANES[J]. Statistical Methods in Medical Research，1999，8(1)：17～36

2　Barnard J，Rubin D B. Small sample degrees of freedom with multiple imputation[J]. Biometrika，1999，86(4)：948～955

3　Fienberg S E，Willenborg L C. Introduction to the special issue：disclosure limitation methods for protecting the confidentiality of sta-

tistical data[J]. Journal of Official Statistics, 1998,14(4):337～345

4 Horton N J, Laird N M. Maximum likelihood analysis of generalized linear models with missing covariates [J]. Statistical Methods in Medical Research, 1999,8(1):37～50

5 King G, Honaker J, Joseph A, et al. Analyzing incomplete political science data: an alternative algorithm for multiple imputation [J]. American Political Science Review, 2001,95(1):46～69

6 Meng X L, Rubin D B. Performing likelihood ration tests with multiple-imputed data sets[J]. Biometrika, 1992,79(1):103～111

7 Robins J M, Wang N. Inference for imputation estimators [J]. Biometrika, 2000,87(1):113～124

8 Rubin D B. Multiple imputation: a primer[J]. Statistical Methods in Medical Research, 1999,8(1):3～15

9 Wang N, Robins J M. Large-sample theory for parametric multiple imputation procedures[J]. Biometrika, 1998,85(4):935～948

10 James H, Anne J, Gary K, et al. Amelia: A Program for Missing Data[S]. Department of Government, Harvard University, 1999. 2～15

11 Aidan M. Multiple imputation for missing data using the "Solas for the missing data analysis" software application[S]. Conference of European Statisticians. 1999. 2～8

12 Nicholas J H, Stuart R L. Multiple imputation in Practices: comparison of software packages for the regression models with missing variables[J]. The American Statistician, 2001,55(3):244～254

13 Yang C Y. Multiple imputation for missing data: concepts and new development[S]. SAS Institute Inc. 1999. 4～9

(收稿日期:2002-09-12)

20种住院死亡疾病年龄统计

赵 炜[1] 胡 军[1] 马 虹[1] 万晓阳[2]

【关键词】 死亡 年龄 统计

中图分类号:R195.4 **文献标识码**:B **文章编号**:1006-5253(2003)02-0081-01

为加强疾病的防治工作,对我院1971—1998年死亡病例进行回顾性统计调查和分析,重点显示严重威胁人类生命的、病死率高的前20种疾病及高发年龄段。

1 资料来源

某院1971—1998年死亡病例原始登记,共4803例。其中超过43例的有20种疾病,共3176例,占死亡总例数的66.1%,男2034例,女1142例。

2 分析

死亡顺位前20种疾病的年龄分布见表1。

表1 20种疾病死亡病例年龄构成比/%

疾病名称	例数	0岁～	11岁～	21岁～	31岁～	41岁～	51岁～	61岁～	71岁～	81岁～
肾炎、尿毒症	178	0.6	8.4	24.2	14.6	18.0	16.9	14.6	1.7	1.1
白血病	128	6.3	15.6	27.3	14.1	14.1	13.3	7.0	2.3	—
脑外伤	111	2.7	6.3	16.2	13.5	15.3	15.3	17.1	9.0	4.5
风心病	96	—	3.1	7.3	17.8	29.2	23.9	14.6	4.1	—
重症肝炎	86	2.3	2.3	10.5	15.1	23.3	23.3	15.1	8.1	—
脑 瘤	69	11.6	4.3	7.2	14.5	20.3	23.2	11.6	5.4	1.4
肝硬化	453	—	1.3	4.0	9.5	28.9	34.7	18.3	3.3	—
肝 癌	272	1.5	0.4	0.7	8.5	20.2	33.5	24.6	9.2	1.5
胃 癌	156	0.6	0.6	1.3	4.5	18.0	25.0	30.0	16.7	3.2
脑外伤	111	2.7	6.3	16.2	13.5	15.3	15.3	17.1	9.0	4.5
肠 癌	62	—	—	3.2	9.7	24.2	19.4	29.0	8.1	6.5
脑出血	375	—	—	1.3	2.1	12.0	27.5	30.1	20.8	6.1
肺 癌	303	0.3	0.3	2.6	5.6	8.9	27.7	33.7	17.5	3.3
肺心病	296	—	—	2.0	2.4	8.1	21.6	32.8	24.3	8.8
脑血栓	144	—	—	0.7	0.7	6.3	28.5	31.9	22.9	9.0
心肌梗死	91	—	—	—	1.1	1.1	25.3	37.4	28.6	6.6
糖尿病	72	—	1.4	9.7	6.9	6.9	25.0	30.6	13.9	5.6
肺 炎	43	—	4.7	2.3	2.3	2.3	18.6	30.2	23.3	16.3
冠心病	95	—	—	—	1.1	3.2	16.8	25.3	27.4	26.3
肺内感染	82	1.2	1.2	1.2	1.2	4.9	12.2	25.6	30.5	22.0
小儿肺炎	64	100.0								

作者单位:1 110003 解放军第202医院 沈阳市
2 沈阳武警医院

(收稿日期:2002-09-03)