【综述】

数据缺失及其填补方法综述

岳勇，田考聪

[关键词] 数据缺失；社会调查；填补方法
[中图分类号] O213 [文献标识码] A [文章编号] 1006－4028(2005)06－683－03

在社会调查资料中，最为常见的问题就是数据缺失。造成数据缺失的原因有：失访、无响应或是回答问题不合格等等。统计学上，将含有缺失数据的记录称为不完全观测。缺失数据或不完全观测对调查研究的影响是很大的。所以在统计学中，为了能够更加充分地利用已经搜集到的数据，国内外很多学者都对缺失数据的处理提出了自己独到的见解，来挽救有缺失的调查数据，以保证研究工作顺利进行[1]。

1 数据缺失概念

缺失的数据是指由于各种原因应该得到而没有得到的数据。在调查研究中缺失数据也被称之为无回答(nonresponse)。统计调查中能否按设计要求获得被调查单位的详全资料是衡量数据质量的一个重要标准，但实际调查中经常遇到数据缺失的情况。

无回答有 2 种表现形式：单位无回答(unit nonresponse)和项目无回答(item nonresponse)。“单位无回答”是指被调查者不原意或者不能够回答整张的问卷；而“项目无回答”则是被调查者拒绝回答个别的调查项目[2]。

2 数据缺失的机制

Little 和 Rubin[2]针对缺失数据定义了 3 种不同的缺失机制。

完全随机缺失(Missing Completely At Random MCAR)：数据的缺失与不完全变量以及完全变量都是无关的。

随机缺失(Missing At Random MAR)：数据的缺失仅仅依赖于完全变量。

不可忽略的缺失(Nonignorable)：不完全变量中数据的缺失依赖于不完全变量本身，这种缺失是不可忽略的。

在实际运用中，要满足 MCAR 情况的假设通常是很困难的，MAR 是常用的假设。

3 缺失数据的处理方法

关于缺失数据在国内外研究的比较系统而深入，其处理方法的研究更是比较新兴的领域，综合国内外现有的研究成果，主要有以下几种方法。

3.1 基于完整观测单位的方法 完整观测单位是指全部调查项目均有观测的单位，或是指在与分析目的相关的调查项目上没有出现“无回答”的单位。该方法就是在剔除数据集中有缺失的单位后，进行常规的统计分析[3]。

3.1.1 删除法(Deletion) 常用的剔除方法有列表删除或个案删除(List wise or case wise data deletion)，配对删除(Pair wise data deletion)。这种方法简单易行，在被调查对象出现多个变量的缺失，并且被删除的含缺失的数据量在整个数据集中的数据量占的比例非常小的情况下，是非常简单而有效的。然而，这种方法却有很大的局限性。它是以减少原始数据来换取数据集信息的完备，会造成资源的大量浪费，丢弃大量隐藏在被剔除对象中的信息。如果数据集中本来包含的对象很少，删除少量对象就足以严重影响到数据集信息的客观性以及结果的正确性；另外在每个变量缺失的百分比变化很大的情况下，它的性能非常差。因此，当缺失数据所占比例较大，特别当缺失数据非随机分布时，这种方法可能导致数据发生偏离，从而引出错误的结论[4]。

3.1.2 加权调整法(Weighting) 加权调整法是考虑对完整观测单位分析进行修正，为尽可能减少分析中可能出现的偏差，从而对完整观测单位进行不同的加权。主要有均值的加权类估计、倾向性加权以及利用加权的广义估计方程进行加权等等。加权是

重庆医科大学卫生统计教研室 (重庆 400016)
作者简介：岳勇(1979－)，男，硕士，主要研究方向：多元统计建模。

一个减少偏差的比较简单的措施，但是由于丢弃不完整单位的信息，并且没有提供一个内在的方差控制，所以在样本量较大时，易出现错误的结果[2]。

3.2 基于填补的方法(Imputation) 该方法的基本思想是利用辅助信息，为每个缺失值寻找替代值。填补法主要用于项目无回答情况。根据所构造的替代值的个数，可以分为单一填补和多重填补[5]。

3.2.1 单一填补法(Single Imputation) 给每一个缺失值构造一个替代值，再对填补后的数据集使用针对完整数据集分析的方法进行分析。常用的方法有：

均值填补法 是指用研究变量回答单位的样本均值作为无回答项目的填补值，分为总均值填补和分层均值填补。前者就是将所有回答单位的均值作为填补值，后者是将样本分为若干填补层后，将各种研究变量的均值分别作为各层所有无回答的填补值。

回归填补法 一种条件性的均值填补法，比一般的均值替代法较为进步。基于完整的数据集，建立回归方程(模型)。对于包含缺失值的对象，将已知变量值代入方程来估计缺失的变量值，以此估计值来进行填补。当变量不是线性相关或预测变量高度相关时会导致有偏差的估计[6]。

Hot Dec 填补法 也叫热平台填补或就近补齐，指在数据集的已知观测中，采用与有缺失的观测最“相似”的那条观测的相应的变量值作为其填补值。这是一种历史比较悠久的填补法，美国普查局多年来都采用了这种方法。它是优于列表删除，配对删除和均数填补的一种缺失数据处理方法。Hot dec 的优点在于概念上简单易懂，可以保持变量本身的数据类型。其缺点在于其中的“相似”很难界定，可以有很多种方法来界定，而且在大型数据集中，运用此方法就显得过于繁琐，并且在模拟数据的分布特征的时候也可能缺乏准确性[7]。

冷平台填补法 用一个从其他来源的常数值代替某一项目的缺失值，比如用以往相类似的调查中的某个值，这样得到的“完整样本”显然是有缺陷的。

以上所述的缺失数据的填补方法中，最主要的问题就是，由于填补的数据都只是唯一的，所以经过填补后的数据集不能表现出原有数据集的不确定性，因而会造成较大的偏差。

3.2.2 多重填补法(Multiple Imputation) 多重填补法是由 Rubin[8]首先提出，经过 Meng[9]和 Schafer[10]等人不断的完善和综合已形成一个比较系统的理论，该法有以下优点：①多重插补过程产生多个中间插补值，可以利用插补值之间的变异反映无回答的不确定性，包括无回答原因已知情况下抽样的变异性和无回答原因不确定造成的变异性。②多重插补通过模拟缺失数据的分布，较好地保持变量之间的关系。③多重插补能给出衡量估计结果不确定性的大量信息，单一插补给出的估计结果则较为简单[11]。

3.2.2.1 具体的多重填补法研究的比较多的有以下几种：① PMM 法 (Predictive Mean Matching, PMM)：也叫随机回归填补法，在回归填补值的基础上再加上残差项，用残差项来反映所预测的值的不确定性。残差项的分布可以是正态分布或非正态分布。这种方法可以保证在正态性假设不成立的情况下，填补适当的值。但难点是随机误差项的确定通常是比较困难。②趋势得分法(Propensity Score, PS)：趋势得分是在给定观测协变量时分配给一个特殊处理的条件概率[12]。在趋势得分法中，对每个有缺失值的变量产生一个趋势得分以表示观测缺失的概率。然后，根据这些趋势得分，将观测分组，再对每一组应用近似贝叶斯 Bootstap 填补[13]。③马尔科夫链蒙特卡罗法(Markov Chain Monte Carlo, MCMC)：MCMC 是贝叶斯推断中的一种探索后验分布的方法。该方法通过填补及后验两步的循环进行，为数据集中的缺失值抽取填补值。

3.2.2.2 多重填补后的综合统计推断 无论采用哪种填补方法，都是将数据集填 m 次(m＞1)，产生个完整的数据集，然后用针对完整数据集的方法来对他们进行分析，并将各个结果加以综合[14]。但到目前为止已经有人研究出针对线性和 logistic 回归模型的综合统计推断法，另外一些如因子分析，结构方程模型和多元正态 logistic 回归模型等等正在研究中[15]。

多重填补出现之初，由于其处理过程比较复杂，在当时并没有得到广泛的应用。在上世纪 80 年代，填补缺失数据的方法也只限于极大似然估计和期望最大化法则(Expectation Maximization EM)算法。直至 90 年代，随着新的计算方法(如 MCMC) 和统计软件的出现，多重插补逐渐成为处理缺失数据的主要工具[16]，应用范围由过去仅能处理项目无回答，发展到既能处理项目无回答，也能处理单位无回答。

3.3 常用的缺失数据填补软件[17] 数据填补通常是一件非常繁琐的工作，很多常用的统计软件以及专门为其编写的软件都可以完成。下面介绍一下常用的针对缺失数据进行填补的软件。

表 1 数据缺失填补软件

软件名称	采用填补方法	假设数据缺失机制	评 价
Amelia[18]	多重填补	MAR	适用于中等专业人员
SPSS	均数填补，期望最大化法则	MCAR，MAR	简单易用
AMOS	原始最大似然法	MAR	简单易用，参数估计，标准误和全局性检验结果可靠
MX	原始最大似然法	MAR	不易使用
NORM	多重填补	MAR	中等难度，但有详尽的帮助
SOLAS[19]	多重填补，Hot Deck，回归	MAR，MCAR	菜单操作，界面简单易用
SAS[20]	均数填补，多重填补，期望最大化法则，混合模型填补	MAR，MCAR	不适用于新手，完全掌握使用很困难

从表 1 中看出，各种软件的着重点不同，使用效果也不尽相同，在使用时要根据个人实际需要加以选择[21]。

综上所述，对于数据缺失国内外学者已经做了广泛的研究。目前各种新兴的方法层出不穷，如人工神经网络，机器智能模型[22]等。但无论采用何种填补方法，都无法避免主观因素对原系统的影响，并且在缺失值过多的情形下将整个数据集完整化是不可行的。就如最早系统提出填补方法的 Rubin 所说："填补，这个概念是十分诱人的也是非常危险的。之所以诱人是它会使人们进入一种可高兴的状态，以至于最后完全迷信填补后的数据集而容易忽略偏差的存在，这即是其危险所在。"[23]所以针对各种实际问题，要注意分清问题的实质，合理并且适当地运用处理方法才是解决好实际问题的关键所在。

4 参考文献

[1] Rubin D. Inference and missing data [J]. Biometrika, 1976, 63(3): 581—592.

[2] Little RJA, Rubin DB. Statistical Analysis with Missing Data [M]. New York: Wiley and Sons, Inc. 1987.

[3] Nordheim EV. Inference from nonrandomly missing data: An example from a genetic study on Turner's Syndrome [J]. Am Statist Assoc, 1984, 79: 772—780.

[4] Horton NJ, Laird NM. Maximum likehood analysis of generalized linear models with missing covariates [J]. Statist Meth Med Res, 1988, 8(1): 37—50.

[5] Allison PD. Multiple imputation for missing data: A cautionary tale [J]. Sociological Methods and Research, 2000, 28(3): 301—309.

[6] Bello AL. Imputation techniques in regression analysis: Looking closely at their implementation [J]. Computational Statistics and Data Analysis, 1995, 20: 45—57.

[7] Rao JNK, Shao J. Jackknife variance estimation with survey data under hot deck imputation [J]. Biometrika, 1992, 79: 811—822.

[8] Rubin DB. Multiple imputations in sample surveys [J]. Am Statist Assoc, 1978: 20—34.

[9] Meng XL, Rubin DB. Performing likelihood ration tests with multiple imputed data sets [J]. Biometrika, 1992, 79 (1): 103—111.

[10] Schafer JL. Analysis of incomplete multivariate data [M]. Chapman and Hall, 1997: 286—293.

[11] Faris PD, Ghali WA, Brant R, *et al*. Multiple imputations versus data enhancement for dealing with missing data in observational health care outcome analysis [J]. J of Clinical Epidemiology, 2002, 55: 184—191.

[12] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects [J]. Biometrika, 1983, 70: 41—55.

[13] Lavori PW. Dawson R, Shera DA. multiple imputation strategy for clinical trials with truncation of patient data [J]. Stat Med, 1995, 14: 1913—1925.

[14] Robins JM, Wang N. Inference for imputation estimators [J]. Biometrika, 2000, 87(1): 113—124.

[15] Barnard J, Rubin DB. Small—sample degrees of freedom with multiple imputation [J]. Biometrika, 1999, 86: 949—955.

[16] Schafer JL. Multiple Imputation: A Primer [J]. Statistical Methods in Medical Research, 1999, 8: 3—15.

[17] Hox JJ. A review of current software for handling missing data [J]. Kwantitatieve Methoden, 1999, 62, 123—140.

[18] James H, Anne J, Gary K, *et al*. Amelia: A Program for Missing Data [J]. Statistical department of Government, Harvard University, 1999: 2—15.

[19] Aidan M. Multiple imputation for missing data using the "Solas for the missing data analysis" software application, [C]. Conference European Statisticians, 1999: 2—8.

[20] Yang CY. Multiple imputation for missing data: concepts and new development [M]. SAS Institute Technical Report, 1999: 1—5.

[21] Nicholas J, Horton and Stuart R, Lispsitz. Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables [J]. The American Statistician, 2001, 55: 3.

[22] Amer, Safaa R. Neural network imputation: A new fashion or a good tool [D]. Oregon State University Statistical Department, 2004.

[23] Rubin DB. Multiple Imputation After 18+ Years [J]. Jof the Am Statist Assoc, 1996, 91: 473—489.

(**收稿日期**：2005—08—29)