基于R语言的缺失值填补方法

李　璐
（暨南大学 统计学系，广州 510632）

摘　要：数据缺失是一个在实验研究和调查研究中经常遇到的问题。文章先介绍了数据缺失机制的四种形式，指出解决数据缺失的一般性方法，即可以通过尽量引入更多的相关变量从而简化缺失机制；然后利用R语言对2006年中国健康与营养调查的部分数据进行了填补，介绍了各种填补方法在R中的应用，并在介绍热平台方法时提出运用R寻找匹配样本的新思路。

关键词：R语言；缺失值；填补方法；缺失机制
中图分类号：C821　　**文献标识码**：A　　**文章编号**：1002-6487(2012)17-0072-03

0　引言

我们在得到一份数据文件时常常会发现文件中存在一些缺失的数据，而缺失数据会对分析任务产生阻碍，造成结果的偏倚及统计工作的低效率。在社会经济调查领域，这些缺失数据的来源是多方面的，包括失访、无回答、录入错误、问题回答不合格等等。对于这部分缺失数据的处理将对分析结果造成一定影响，而现实情况是分析数据者往往对于缺失数据的处理方法没有相应地专业知识，因此轻易对缺失数据进行删除或简单的填补，而没有考虑到更深层的因素。本文拟在介绍数据缺失机制及相关理论的基础上，以2006年中国健康与营养调查的部分数据为例，介绍各种填补方法在R语言中的应用，并在介绍热平台方法时提出运用R寻找匹配样本的新思路。

1　缺失数据的理论知识

1.1　数据缺失机制

处理缺失数据前，首先应该了解数据缺失的原因，也就是缺失机制[1]。缺失机制是指缺失变量与分析变量的关系，了解数据缺失的原因，有利于选择合适的处理方法对数据进行处理。一般情况下，缺失机制可以分为以下四种类型[2]：

(1)完全随机缺失，假如缺失的概率对于各变量的取值是等概率的，即缺失是完全随即的，那么删除缺失数据后的结果将是无偏的。

(2)随机缺失，指缺失的概率只与数据集中被观察到的值有关，与未观察到的值无关，通常情况下，完全随机缺失的假设很难被满足，而随机缺失则是一个相对宽松的假设。

(3)基于未观测变量的缺失，更常见的情况是，缺失的概率不仅与已观测变量有关，可能还与未观测到的变量有关。

(4)基于缺失值本身的缺失，指缺失的概率依赖于缺失值本身，一般可以认为生存分析中的删失数据属于这类缺失机制，本文暂不考虑这类缺失。

知道了数据的缺失机制后，便能对数据进行初步处理。对于完全随机缺失的数据，简单地删除并不会对结果造成偏倚。对于随机缺失的数据，也是易于处理的，可以通过建立缺失变量与影响缺失概率的变量之间的回归模型进行预测。而对于基于未观测变量的缺失，以及基于缺失值本身的缺失这两种缺失机制，则要复杂得多，原因是我们既然无法观测到相关变量的值，也就无法判断它们之间的内在关系。解决这个问题的一个方法是尽量将所有类别的缺失机制都简化到随机缺失这一类缺失机制。由于完全随机缺失是少见的，并且将完全随机缺失作为随机缺失考虑依然不影响结果的偏倚性。而将基于未观测变量的缺失简化为随机缺失的想法是可以将尽量多的变量作为相关变量加入到模型中，这些变量可能与未观测到的相关变量存在相关关系，可以将其看作是工具变量。因此，本文所使用的方法就有了理论基础。

1.2　缺失数据的处理方法

1.2.1　删除法

解决缺失数据问题的一个简单易行的方法是删除部分数据，使之成为完整数据进行分析，而这些程序在R语言中是简单实施的，根据分析的角度不同，删除数据的方法可以分为四种类型：(1)观测样本删除。即将存在缺失值的样本直接删除，这是最直接的删除数据的方法。在R语言中，做回归分析时，系统会将含有缺失值的样本剔除在外。(2)变量删除。当某个变量的缺失率较大时，比如某变量有一半的调查者无回答，可能的原因是问卷设置的问题，可将该变量删除。(3)完全变量分析。即在试验研究中，有时候研究的某一个问题可能只涉及到某几个变量，

作者简介：李　璐(1988-)，女，湖南岳阳人，硕士研究生，研究方向：统计预测与决策。

而这几个变量的原始数据是完整的，那就可以只分析这几个完整的原始数据，而无需使用缺失的不相关数据。(4)无回答权重。当缺失值类型为非完全随机缺失的时候，可以通过对完整的数据加权来减小偏差。把数据不完全的个案标记后，将完整的数据个案赋予不同的权重，个案的权重可以通过logistic回归求得[3]。

1.2.2 填补法

在实验研究中，删除数据可能被认为是信息量与研究花费的巨大浪费，为避免这种情况，更好的方法是对缺失值进行填补，用填补值尽可能接近真实值来还原数据。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。根据所构造的填补值个数，可以分为单一填补和多重填补。本文重点介绍单一填补，包括均值填补，回归填补，二阶填补，热平台，冷平台，抽样填补等方法，多变量情形可按照单变量情形推及。

2 缺失数据的填补方法[4]

本文示例数据选用2006年中国健康与营养调查的成人问卷数据，并筛选部分数据作为分析数据，其中使用的变量包括年龄(A3A)、性别(AA2A)、受教育程度(A12)、工作单位类型(B6A)、每天工作时间(C6)、工资(C8)、是否干农活(D2A)、是否照顾儿童(K12)、吸烟(U25)、睡眠时间(U324)、是否上网(U354)。本文先考虑单变量情形，即工资变量(C8)，使用的样本量为1857个，再推广为多变量情形。

2.1 数据特征

```
> table(cut(C8,breaks=c(0,100,1000,5000,10000,100000)),exclude=NULL
```

表1 工资数据频数表

工资(C8)	(0,100]	(100,1000]	(1000,3000]	(3000,10000]	(10000,100000]	NA
频数	6	1135	633	23	16	45

从表1可以看出工资数据缺失45个，并且当月工资高于3000是已经非常少，为了方便表现数据的特征，将高于3000的部分数据设定为3000。

2.2 单变量填补

2.2.1 简单随机填补

从抽样的角度考虑，最简单的方法是从已有的工资变量的样本数据直接用随机抽样抽取样本值作为填补值进行填补。R程序如下：

```
>random.imp<-function(x)
{ missing=is.na(x)
 n.missing=sum(missing)
 x.obs=x[!missing]
 imputed=x
 imputed[missing]=sample(x.obs,n.missing,replace=T)}
>simp.ran.c8=topcode(c(random.imp(C8),C8),3000)
```

从图1可以看出，随机抽样抽取的填补数据与原数据的特征相似。这种方法虽然从理论到运用都简单易懂，但是，很明显的一个缺陷是这种方法仅仅考虑到了缺失变量本身，而并没有考虑到相关变量的信息。因此，信息量的利用少，效果不理想。

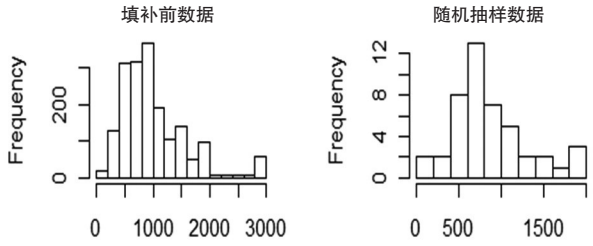

图1 抽样数据与原数据进行比较

2.2.2 均值填补

与随机抽样填补法一样，仅利用缺失变量数据的另一种方法是均值填补，即计算缺失变量的均值，用均值作为填补值。当然，类似用均值填补这种思路，也能用中位数填补，1/4分位数和3/4分位数的均值填补等等，更复杂的，可以考虑这些填补方法的偏倚值来选取适当的方法。其中，均值填补的R程序如下：

```
>m.imp=topcode(c(rep(mean(C8),45),C8),3000)
```

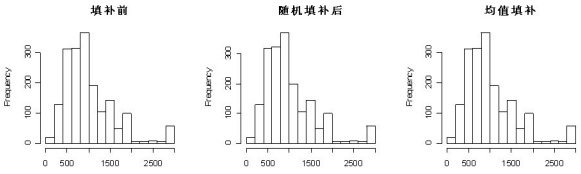

图2 随机插补和均值插补前后对比

从图2可以看出，均值填补的效果和随机填补相似。另外，本文所使用的例子的数据不是太好，原因是缺失的样本数量相对于总体样本的数量很小，因此很难看出填补数据后与填补数据前的差别。

2.2.3 回归填补

区别于上诉两种忽略相关变量信息的方法，直观上更有效的方法是拟合一个回归，将缺失变量作为因变量，相关变量作为自变量，以预测值作为填补值。考虑拟合一个简单线性回归模型，R中会自动忽略具有缺失值的样本，因此选择的变量应该具有完全数据，程序如下：

```
>lm.imp=lm(C8~A3A+AA2A+B6A+D2A+K12+U25+U354,data=A,subset=C8<3000)
 summary(lm.imp)
 pred=predict(lm.imp,A)
>impute<-function(x,x.impute)
 {
 ifelse(is.na(x),x.impute,x)
 }
>C8.imp=impute(C8,pred)
```

表2 回归模型

模型	Adjusted R-squared	F-statistic	p-value
lm.imp	0.1302	38.51	2.2e-16

从表2看出，拟合并不理想，为此可以加入更多的自变量。一个改进的思路是改善自变量的数据质量，即对有缺失值的变量先进行填补，使之成为完全数据，这样该变

量就能包含进来。另一个改进的思路是对预测变量进行变换,包括开方,对数化,求倒数等等,如对工资变量开平方,使数据接近正态分布。

2.2.4 热平台和冷平台

匹配插补又称热平台方法,对一个具有缺失值的样本,我们在具有完全数据的样本里面寻找一个与它相似的样本,用该样本的值填补缺失值。相对应的,冷平台方法[5],又称条件均值插补法,是指根据相关变量将总体分层,对于任一缺失值,用该样本所在层的完全数据的均值代替。

R中寻找匹配样本的一个思路是利用上面拟合的模型lm.imp,计算各样本缺失变量的预测值,将相邻最近的缺失数据样本和完全数据样本作为匹配样本,用完全数据样本的值填补该缺失值。

2.2.5 随机回归填补法

结合随机抽样和回归法的方法是随机回归填补法,利用上面的回归模型lm.imp所得的预测值进行模拟,程序如下。从填补的效果来看,填补值出现了负数,原因可能是数据不符合正态分布。在分布未知的情况的,更好的方法是用自助法(Bootstrap)进行模拟[6]。

2.3 多变量填补

上诉内容仅仅讨论了单变量缺失的情形,通常情况下,我们得到的数据常有多于一个变量是具有缺失值的,这时候,上面的方法就可能不是那么适用了。

从上面单变量填补法进行推广,解决多变量缺失的一个直接的方法是回归插补法,通过对缺失变量与完全数据变量拟合多元回归模型来预测缺失值。这一方法的具体应用是多重填补法[7],而上述的随机回归填补是多重填补的其中一种。

3 总结

本文介绍了R中对缺失值填补的一些方法,包括简单随机填补,均值填补,回归填补,热平台和冷平台,随机回归填补,并针对这些方法提出了一些可以改进的新思路。另外,本文也有不少不足。首先本文适用例子的缺失值数只有总数据量的2.5%,因此很难看出填补的效果;其次对多变量缺失的讨论与研究尚未进入,这可作为以后的一个研究方向。

参考文献:

[1]Huisman M,Krol B,Sonderen EV. Handling Missing Data by Re-Approaching Non-Respondent[J].Quality & Quantity,1998,32(1).

[2]Graham JW,Donaldson SI. Evaluating Interventions with Differential Attrition: The Importance of Nonresponse Mechanisms and Follow-up Data[J]. Applied Psychology,1993,78(1).

[3]胡红晓,谢佳,韩冰.缺失值处理方法比较研究[J].商场现代化,2007,(504).

[4]Andrew Gelman,Jennifer Hill. Data Analysis Using Regression and Multilevel/Hierarchical Models[M]. Cambridge:Cambridge University Press,2007.

[5]Gadbury GL,Coffey CS,Allison DB.Modern Statistical Methods for Handing Missing Repeated Measurements in Obesity Trial Data:beyond LOCF[J].Obesity Rev.,2003,(4).

[6]Sheldon M.Ross.Simulation统计模拟[M].王兆军,陈广雷,邹长亮 译.北京:人民邮电出版社,2007.

[7]袁中荧.多元线性回归模型中缺失数据填补方法的效果比较[D].中南大学硕士学位论文,2008.

[8]毛群霞.缺失值处理统计方法的模拟比较研究及应用[D].四川大学硕士学位论文,2005.

[9]王斌会.R语言统计分析软件教程[M].北京:中国教育文化出版社,2007.

(责任编辑/浩 天)