第四章

统计数据的缺失值处理

数据分析时经常遇到数据缺失的情况，缺失数据已成为影响数据质量的一个重要因素。现实中导致数据缺失的原因较多，主要来自于调查中的无回答、遗漏调查项、不合理数据剔除等因素。缺失数据的存在往往导致明显的非抽样误差，造成估计量方差增大，很多情况下还会引起系统性偏差。所以，处理好缺失数据，对于提高统计数据质量具有十分重要的意义。

目前我国有关缺失数据处理的文献并不多，但国外对这一问题已进行了广泛的研究。金勇进等(2009)将关于统计调查中缺失数据问题的理论研究划分为三个阶段[①]：第一阶段是宣传期(1915 年－20 世纪 40 年代)，有关学者开始对缺失数据问题的初步研究，强调处理缺失数据问题的重要性。第二阶段是专题研究、方法发展期(20 世纪 40 年代中期－90 年代初)，许多学者对缺失数据问题进行了大量的专题研究，提出对缺失数据进行补救的经典方法。第三阶段是方法完善期(20 世纪 90 年代初至今)，该时期很少有学者提出无回答处理的全新思想，但许多学者进行了经典方法的改进和拓展，开展理论总结、方法比较，并强调实际应用。以下借鉴现有研究成果，依据缺失数据的机制和模式，对缺失数据的事前和事后处理方法及其应

① 金勇进、邵军：《缺失数据的统计处理》，中国统计出版社 2009 年版。

用做一个综合性探讨。

一、缺失数据的机制和模式

(一)缺失数据产生的机制

缺失数据的产生机制是通过探讨缺失数据的出现与目标变量是否有关而界定的,如果缺失数据的出现是随机的,则将该类缺失数据的产生机制定义为可忽略的;如果缺失数据的产生与研究变量有关,称之为不可忽略的[①]。

定义完全数据 $Y=(y_{ij})$,缺失数据矩阵为 $M=(M_{ij})$,缺失数据的机制由给定 Y 时 M 的条件分布来刻画,记为 $f(M|Y,\phi)$,其中 ϕ 是未知参数。如果缺失不依赖数据 Y 的值,也就是如果

$$f(M|Y,\phi)=f(M|\phi) \quad (\text{对所有的 } Y,\phi) \tag{4-1}$$

则数据称为完全随机缺失(MCAR)。令 Y_{obs} 为 Y 的已观测部分,Y_{mis} 为缺失部分,若缺失仅依赖 Y 已观测的部分 Y_{obs},不依赖缺失部分,即

$$f(M|Y,\phi)=f(M|Y_{obs},\phi) \quad (\text{对一切 } Y_{mis},\phi) \tag{4-2}$$

则这种缺失机制称为随机缺失(MAR)。如果 M 的分布依赖数据阵 Y 的缺失值,这样的机制称为非随机缺失(NMAR)。可以看出缺失数据产生机制主要是关注 Y 给定条件下 M 的分布。

金勇进等(2009)认为缺失数据的产生机制除以上三类之外,还可分为取决于协变量的缺失(CDM)、取决于随机影响缺失(REDM)和取决于前期数据的缺失等几种[②]。取决于协变量的缺失是指目标变量 y 是否缺失与协变量(辅助变量)x 有关,而与缺失的 y 值无关;取决于随机影响缺失和取决于前期数据缺失经常存在于纵向数据研究中。

总体来看,在可忽略机制之下产生的缺失数据处理相对容易,处理方法容易掌握,而在不可忽略机制之下产生的缺失数据处理比较

---

① Roderich J. A. Little and Donald B. Rubin. 孙山泽译:《缺失数据统计分析(中文版)》,中国统计出版社 2004 年版。

② 金勇进、邵军:《缺失数据的统计处理》,中国统计出版社 2009 年版。

困难，原因在于偏差的程度难以把握。

（二）数据缺失的模式

如果在众多观测变量中缺失数据仅限于单个变量，其他变量的值都能够观测得到，此类缺失数据称为单变量缺失模式；与单变量缺失模式相对应的是多变量缺失模式，即在观测过程中存在多个变量值缺失的情况。

当一维目标变量出现缺失数据时，在数据预处理过程中主要考虑缺失数据产生的机制；而对于多维目标变量而言，除了考虑缺失数据产生的机制外，还需判断数据的缺失模式①。

从缺失的对象来看，可分为个案缺失和项目缺失两种情况。个案缺失是指选入样本的单元没有给出回答，原因可能是调查时被访者不在家，或者拒访，或者由于某些原因无法接受调查等。另一类为项目缺失，指被访者接受调查，回答了一部分问题，而有些问题没有数据，可能是被调查者没有回答这些问题，也可能是调查人员漏问、漏记，也包括对离群数据的删除等。

另外，从缺失数据表现形式的角度还可将缺失数据分为单调缺失模式和任意缺失模式。假设完全数据资料 Y 是由 m 个观测、n 个变量组成的 $m \times n$ 矩阵，对这个矩阵进行适当的行列变换后，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 Y_{ij} 缺失时，则对任意的 $p \geqslant i$ 和 $q \geqslant j$，元素 Y_{pq} 也是缺失的，这种数据缺失模式被称为单调缺失模式（monotone missingness pattern）。不满足单调缺失模式的，被称为任意缺失模式（arbitrary missingness pattern）。

对于单调缺失模式来说，缺失数据的处理比较简单，但在大多数复杂的调查中，这种缺失模式很少见。对任意缺失模式而言，处理方法较为复杂。如果可能的话，可以先将非单调缺失资料变换为单调缺失，之后再采用针对单调缺失模式的处理方法。

可见，针对缺失数据产生的不同机制及表现出的不同模式，对其

① 庞新生：《缺失数据处理中相关问题的探讨》，《统计与信息论坛》2004 年第 5 期。

进行预处理的方法也存在较大差异，但总体可分为事前处理方法和事后处理方法两大类。

二、缺失数据的事前处理方法

（一）事前预防措施

从理论上讲，事前预防也许是处理缺失数据最简便且最有效的方法。在对缺失数据问题的研究中，早期学者也较多地关注了缺失数据的事前预防方法和措施。Kish(1965)、Warwick-Lininger(1975)、Mosteller(1978)等都对提高缺失数据回答率的措施进行过广泛的讨论。Deming(1953)、Dubin(1954)以及后来的 Thomsen 和 Siring(1983)采用不同的方法来决定访问调查中理想的尝试次数。Dohrenwend(1970)、Ferber 和 Sudman(1974)、Chromyt 和 Horvitz(1978)、Gunn 和 Rhodes(1981)曾研究过激励方法对改善回答率的效果[①]。

数据收集过程中采取一定的事前预防措施，可减少由于无回答、填报和汇总等原因造成的缺失数据。具体包括以下措施：

首先，注重调查表的设计质量，合理安排调查项目，问题由简单到复杂，由总体到局部，并且努力降低问题的敏感性；

其次，加强调查员和数据录入、汇总人员的选拔和培训，增强他（她）们的责任心和业务能力，避免由此而造成的缺失数据；

再次，充分利用有关媒体加强宣传和激励，使被调查者全面了解调查内容及重要性，提高其参与意识；

另外，由于统计调查总要占用被调查者的时间和精力，对那些同意接受调查并提供回答的被调查者应给予适当的物质奖励，可以调查开始前做出说明，增加被调查者接受调查的积极性。

（二）敏感性问题的事前处理

在调查过程中，因无回答造成的数据缺失很大程度上是由于调查所问的问题具有较强的敏感性，被调查者不愿回答，所以在缺失数

① 金勇进、邵军：《缺失数据的统计处理》，中国统计出版社 2009 年版，第 11—15 页。

据事前处理方法中，对敏感性问题的处理需重点加以关注。

敏感性问题是指与个人或单位的隐私或私人利益密切相关而不便向外界透漏的问题，例如是否存在行贿、受贿情况；考生是否存在考试作弊；个体工商户是否偷税、漏税；是否为同性恋者等问题。对于这些敏感性问题，若采用直接提问的形式，被调查者难免产生抵触情绪，不愿回答，容易造成数据缺失。因此，寻求解决敏感性问题调查的有效方法对于减少数据缺失、提高数据质量至关重要。

对敏感性问题的处理，关键是使被调查者既愿意做出回答，又能保守个人秘密。心理学家与统计学家为此设计了一些调查和分析方法——随机化回答技术，其中主要包括沃纳模型、西蒙斯模型，以及在此基础上的改进方法——随机变量和回答模型[①]，此类方法的基本原理通过以下例子进行说明。在调查学生考试作弊的问题中，设计外形完全一样的卡片 n 张，其中 n_1 张卡片上写着“你考试是否作过弊?”，$n-n_1$ 张卡片上写着另外的问题，然后放在同一个盒子里；调查时，由被调查者从盒子里任抽一张卡片，根据卡片上的问题做出回答，至于卡片上具体是什么问题，调查者无权过问，这样就起到了为被调查者保密的作用，易于得到被调查者的合作。

1. 沃纳随机化回答模型

由美国统计学家沃纳提出，为了调查某个敏感问题，同时列出两个存在相关关系的问题并制成卡片，被调查者随机抽取卡片进行回答。具体的做法是：针对需要调查的敏感性问题，列出正反两个问题，譬如调查考试作弊问题，就做成两种卡片：

A. 你在考试中作弊了吗?

B. 你在考试中没有作弊吗?

然后，由被调查者随机抽取一张来回答“是”或“否”，至于抽取的卡片上具体是什么问题，调查者无权过问。如果被调查者所抽到的问题与自己情况一致则回答“是”，否则回答“不是”。因此，调查人员并不知道被调查者在回答哪一个问题，从而达到对被调查者个人隐

① 石艳芬：《敏感性问题调查的基本方法与比较》，《统计与信息论坛》2002 年第 5 期。

私的保密效果。虽说调查人员不知道被调查者具体抽中的是哪一个问题，但问题 A 所占的比例 P 是确定的。

设 Π_A 是具有敏感性特征的人所占比例，p 是写有问题“A”的卡片所占的比例。如果对 n 人进行调查，调查结果中有 n_1 个人回答“是”，有 $n-n_1$ 个人回答“否”，调查结果中回答“是”的人的比例 $\phi = n_1/n$，对问题 A 回答“是”的人数比例为 Π_A。

Π_A 的极大似然估计为：

$$\hat{\pi}_A = \frac{\frac{n_1}{n} - (1-P)}{2P-1} \tag{4-3}$$

方差为：

$$\mathrm{Var}(\hat{\pi}_A) = \frac{\frac{n_1}{n}\left(1-\frac{n_1}{n}\right)}{n(2P-1)^2}$$

2. 西蒙斯模型

沃纳的方法虽然比直接提出敏感性问题要好，但所提的两个问题都还具有一定的敏感性，而且该方法中回答问题“A”的人数比例不能为 1/2。1967 年西蒙斯对沃纳模型进行了改进，得到西蒙斯模型，其与沃纳模型最大的不同点在于调查人员提出的随机化问题是两个不相关的问题，其中一个为敏感性问题，另一个为非敏感性问题，这样的处理使被调查者的合作态度进一步提高。

设样本中对问题“B”（无关问题）回答“是”的人数比例为 Π_B，Π_A 和 ϕ 的含义同沃纳模型，ϕ 为调查结果中回答“是”的人所占比例，也就是对问题“A”或“B”回答“是”的人数比例，Π_A 是对问题“A”回答“是”的人数比例。

(1)Π_B 已知的情况。设抽样方式是简单随机有放回抽样，Π_A 是具有敏感性特征“A”的人所占比例；总体为 n 的简单随机样本中，有 n_1 人回答“是”，则 $\hat{\phi} = n_1/n$。

Π_A 的极大似然估计为：

$$\hat{\pi}_A = \frac{\hat{\phi} - \pi_B(1-P)}{P} \tag{4-4}$$

方差为：

$$\mathrm{Var}(\hat{\pi}_A) = \frac{\phi(1-\phi)}{nP^2}$$

(2)Π_B 未知的情况。此时需要抽取两个随机样本进行调查，设两个样本的容量分别为 n_1 和 n_2，敏感性问题占的比例分别为 P_1 和 P_2。假设总体 1 中回答敏感性问题的人所占比例为 P_1 时，对问题“A”或“B”作出“是”的答复者所占的比例为 ϕ_1；总体 2 中回答敏感性问题的人所占比例为 P_2，对两个问题作出“是”的答复者的比例为 ϕ_2，从而得到该敏感性问题回答“是”所占比例的估计值为：

$$\hat{\pi}_A = \frac{\phi_1(1-P_2) - \phi_2(1-P_1)}{P_1 - P_2} \tag{4-5}$$

方差为：

$$\mathrm{Var}(\hat{\pi}_A) =$$

$$\frac{1}{(P_1-P_2)^2}\left[\frac{\phi_1(1-\phi_1)(1-P_2)^2}{n_1} + \frac{\phi_2(1-\phi_2)(1-P_1)^2}{n_2}\right]$$

3.敏感性问题处理方法的应用

某高校在开展关于普及性知识的活动中，要求对学校的学生“是否有过性行为”这一问题进行抽样调查。该调查问题具有很强的敏感性，如果运用通常的调查方式调查根本无法进行，因此需要运用针对敏感性问题抽样调查方法。该高校在校生人数为 6000 人，随机抽取 1500 名学生进行抽样调查，且分别运用上述两种方法，比较统计结果①。

第一种方法采用沃纳模型：提出两个都具有敏感性的相关问题。

采用随机化回答技术设计两种用信封封装比例且一定的问卷，一种问题为“你有过性行为吗?”另一种问题为“你没有过性行为吗?”调查时，让同学任意选取一个信封并回答上面的问题，当然调查人员不知道该同学回答的是哪一个问题。

同学们根据实际情况回答所抽到的问题，与自己的情况一致的

① 案例资料参考张天哲：《敏感问题的调查技术》，http://www.quanwen.com.cn/doc/2267754/，2009 年。

则回答“是”；否则回答“不是”。研究者在设计问卷时，设计第一种问题占 60%，这样两个问题所占的比例比较接近，有助于让被调查者消除顾虑。对收回的问卷进行统计，发现对两种问题回答“是”的有 638 人，占样本的比例为：$\phi = 638/1500 = 0.4253$。

已知 $\phi = 0.4253$，$P = 60\%$，则回答第一种问题为“是”的人数估计比例：

$$\hat{\pi}_A = \frac{0.4253 - (1 - 0.6)}{2 \times 0.6 - 1} = 0.1265$$

方差为：

$$\mathrm{Var}(\hat{\pi}_A) = \frac{\phi(1-\phi)}{n(2P-1)^2} = \frac{0.4253 \times (1 - 0.4253)}{1500(2 \times 0.6 - 1)^2} = 0.0041$$

第二种方法采用西蒙斯模型：提出的两个问题，一个为敏感性问题，另一个为与调查内容无关的非敏感性问题。

同样采用随机化回答技术设计两种用信封封装且比例一定的问卷，然而一种问题为“你有过性行为吗？”为了方便选择已知的情况，即另一种问题设计为“你是四月份出生的吗？”

显然，第二个问题与所要调查的问题无关，而且被调查同学当中是四月份出生的比例很容易从学校教务处学生信息中心收集到，经查该校学生中四月份出生者所占的比例为 15.38%。设计的问卷中第一种问题同样占 60%，统计结果为对两种问题回答“是”的有 206 人，占样本的比例为：$\phi = 206/1500 = 0.1373$。

得到回答第一种问题为“是”的人数估计比例为：

$$\hat{\pi}_A = \frac{0.1373 - 0.1538(1 - 0.6)}{0.6} = 0.1263$$

方差为：

$$\mathrm{Var}(\hat{\pi}_A) = \frac{\phi(1-\phi)}{nP^2} = \frac{0.1373 \times (1 - 0.1373)}{1500 \times (0.6)^2} = 0.0002$$

三、缺失数据的事后处理方法

现实中由于种种原因和条件的限制，事前处理方法往往并不能使数据缺失的问题得到完全解决。因此，缺失数据的事后补救方法

越来越受到重视，很多学者对此进行了深入的理论和实证研究。缺失数据事后处理方法中的代表性方法是加权调整法，Deming 和 Stephan(1940)提出事后分层重复多变量逐一加权的方式，Hansen 和 Hurwitz(1943)提出按照样本抽取率的倒数加权，Politz 和 Simmons(1949)提出经典的 Politz-Simmons 调整法，按照回答者在相同时间内在家并可接受调查的天数进行加权，Horvitz 和 Thompson(1952)提出按照单位被抽中概率的倒数加权[①]。后期各种推陈出新的加权方法基本上都承袭了早期的这些观念。

加权法主要用于单位缺失数据的补救处理，而对于项目缺失数据的补救处理则多采用插补法(也称为“替代法”或“估算法”)。学者们陆续提出均值插补、热卡插补、冷卡插补、回归插补和模型插补等方法，并进行了广泛讨论和改进。Nordbotten(1963)和 Chapman(1976)探讨了冷卡法在周期性调查的作用；Sonquist(1971)、Chapman(1976)、Ford(1983)、Rizvi(1983)、Sande(1983)等对热卡插补法进行过讨论和改进；Kalton 和 Kish(1984)、Sande(1979，1982)在热卡法的基础上提出数值分类的距离函数匹配法，以避免回归插补和热卡插补法的困境。此外，Hansen 和 Hurwitz(1946)提出以传统的统计推论为基础的双重抽样法，以 Rao(1972)、Singh(1978)等人为代表，探讨了贝叶斯方法在缺失数据处理中的应用。Dempster、Laird 和 Rubin(1977)提出一种有效估计不完全数据算法——EM 算法，不仅是一种有效的计算工具，还根本性地改变了统计学家对缺失数据的看法[②]。正是基于这一算法，Rubin 在 20 世纪 80 年代初期的一系列论文中提出了多重插补法。

(一)事后处理的基本方式

面对缺失数据，所谓事后处理就是认真分析其产生的原因，积极采取有针对性的补救措施。事后处理的基本方式包括以下几种：

---

① 金勇进、邵军：《缺失数据的统计处理》，中国统计出版社 2009 年版，第 11—15 页。

② 金勇进、邵军：《缺失数据的统计处理》，中国统计出版社 2009 年版。

1.忽略缺失数据，直接分析

不考虑缺失数据影响，直接在已获取数据的基础之上进行统计分析。主要包括以下两种情况：(1)删除有缺失值的个案。将包含缺失值的个案直接从分析对象中剔除，利用包含完整信息的个案进行统计分析。如果数据收集过程中控制得不是很好，被访对象多多少少都会出现一些问题没有回答的情况，如果直接删除这些个案必然导致样本量的减少。(2)保留有缺失值的个案，仅在相应的分析中做必要删除。如果样本量比较大，缺失值的数量又不是很多，而且变量之间也不存在高度相关的情况下，研究者经常采用这种方式处理缺失值。该方法比较容易执行，但也容易导致严重的偏差。

2.开展再调查，补充缺失数据

通过多种方式重新抽取样本，对缺失数据进行相应的补充。具体形式包括：(1)多次访问。对数据收集过程中无回答单位进行再一次补充调查，以尽可能多地获得调查数据。如果缺失数据是在不可忽略机制下产生，由于积极回答者和不回答者之间的数量特征有较大差异，多次访问很有必要，且这种差异越大，访问的次数也相应增加。(2)替换被调查单位。在出现无回答的情况下，为了使样本量不低于原设计要求，一个补救的方法是实行替换，用总体中最初未被选入样本的其他单位去替代那些经过努力后仍未获得回答的单位，使用替换法应尽可能保证替代者和被替代者的同质性。(3)对无回答单位进行子抽样。当后继访问的单位费用昂贵时，子抽样被作为减少访问次数的一种现成方法。

3.利用辅助信息，进行间接估计

在进行统计分析时，不仅考虑已获得的数据，而且用到一些相关的辅助信息来改善和调整缺失数据带来的影响。具体包括：(1)加权调整。该方法的基本思想是用调整因子来调整利用包含无回答数据进行的总体推断，即将赋予缺失数据的权数分摊到已获得数据身上。应用该方法的前提是假定缺失数据在可忽略机制下产生，即获得数据与缺失数据之间没有显著差异。在分层抽样条件下，调整可以在各层分别进行。该方法主要用于个案缺失情况下的调整。(2)插补。

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。插补法主要用于项目无回答情况下的调整，根据每个缺失值替代值的个数，可以分为单一插补和多重插补。所谓单一插补是指对每一个由于无回答造成的缺失值只构造一个替代值；而多重插补是指给每个缺失值都构造一个以上的替代值，这样就产生了若干个完全数据集，对每个完全数据集分别使用不同的方法处理，得到若干个处理结果，最后再综合这些处理结果，最终得到目标变量的估计。插补法的效率如何，取决于插补值与原无回答数据的相似程度。

（二）事后处理的具体方法

因缺失数据的存在，往往影响到利用数据进行判断和决策的准确性，所以通常需要利用一定的方法对缺失值进行处理。由于开展再调查补充缺失数据很多时候不太现实，故对于缺失数据的实际事后处理方法主要包括直接删除缺失数据、加权调整法以及插补法，其中插补法是几种方法中应用最为广泛的方法，相应的研究也最多。

1. 个案剔除法

处理缺失数据最常见、最简单的方法是个案剔除法（listwise deletion），也是很多统计软件（如 SPSS 和 SAS）默认的缺失值处理方法。该方法意味着如果任何一个变量含有缺失数据的话，就把相对应的个案从分析中剔除。如果缺失值所占比例比较小，这一方法十分有效，至于具体多大的缺失比例算是“小”比例，专家们意见也存在较大差距。当然，这种方法也有很大的局限性，它是以减少样本量来换取信息的完备，会造成资源的大量浪费，且丢弃了大量隐藏在缺失对象中的信息。在样本量较小的情况下，删除少量对象就可能严重影响数据的客观性和结果的正确性。因此，当缺失数据所占比例较大，特别是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，从而得出错误的结论。

2. 加权调整法

加权调整法是通过对未缺失数据使用加权因子对其进行调整，

减小由于缺失数据造成的估计偏差①。设从总体 N 中随机抽取容量为 n 的样本，估计量 $\hat{Y}=\sum_{i=1}^{n} w_i y_i$ 是无偏的，w_i 是第 i 个样本单位的权数；若令 π_i 为第 i 个单位的入样概率，在样本单位全部回答情况下，权数 $w_i=\pi_i^{-1}$，反映了第 i 个样本单位在估计中的作用。又设 p_i 为第 i 个单位的回答概率，$p=1$ 表示一定回答，$p=0$ 表示一定不回答，现实中 p_i 是一个随机变量，被调查者是否回答取决于多种因素。设回答概率期望值 $E(p_i/\pi_i=1)=p_i$，即第 i 个单位被选中后的回答概率为 p_i。在调查中，由于无回答的存在，只能用 n_1 个回答单位的信息对总体参数进行估计，因此估计量 $\hat{Y}=\sum_{i=1}^{n} w_i y_i$ 就需要修正为 $\hat{Y}^*=\sum_{i=1}^{n_1} w_i^* y_i$，其中 $w_i^*=(\pi_i p_i)^{-1}$ 是对由于无回答造成缺失数据进行调整的权数。

实际上，调整是根据调查中回答单位的回答概率进行的。为进行调整，需要掌握样本单位的回答概率，但现实中 p_i 通常未知，需要对 p_i 进行合理的估计。对 p_i 的不同估计就形成不同的调整方法。因此，加权调整法是一个概括性的说法，包含了一序列不同的调整方法，譬如 Politz-Simmons 调整法、加权组调整法、再抽样调整法、事后分层调整法、迭代调整法、校准法、双重稳健加权法等。

3. 均值插补法

均值插补法(mean imputation) 将变量的属性分为数值型和非数值型来分别进行处理，如果缺失值是数值型，就利用该变量在其他所有对象中取值的平均值来填充该缺失的变量值；如果缺失值是非数值型，就根据统计学中的众数原理，用该变量在其他所有对象中取值次数最多的值来补齐该缺失的变量值。均值插补法是一种简便、快速的缺失数据处理方法，在缺失值是完全随机缺失(MCAR) 时为总体均值或总量提供无偏估计。然而它严重扭曲了数据分布，所有的插

① 金勇进：《缺失数据的加权调整》，《数理统计与管理》2001 年第 5 期。

补值集中在均值点，在分布上形成尖峰，导致方差低估。

（1）无条件均值插补。无条件均值插补是用所有记录单元的均值来插补缺失值。此时得到的总体均值和方差的估计量分别是$\bar{y}_j^{(j)}$和$S_{jj}^{(j)}(n^{(j)}-1)/(n-1)$，其中$\bar{y}_j^{(j)}$表示变量$y_j$的所有记录单元的均值，$S_{jj}^{(j)}$是有记录单元的方差，$n^{(j)}$表示有记录单元的总个数。在MCAR的假定下，总体均值的估计量是无偏估计，方差及和变量y_j的协方差估计量是相合估计，但由于插补值是来自分布中心的数值，因而扭曲了变量的经验分布，此时总体方差和协方差被低估。

（2）条件均值插补。在无条件均值插补中，对于所有缺失数据均用有记录单位的均值进行插补，得到的是过于集中的扭曲经验分布。为了改善这种状况，让插补后的数据更好地反映总体的真实波动，从而得到更加准确的方差、分位数等估计量，有必要进行条件均值插补。① 分层均值插补。在进行插补之前，对变量Y按照数据中的某一个变量分层，然后在每一层中，用该层有记录单元的均值插补该层的缺失值。在MAR的假定下，该方法对于总体均值的估计是无偏的。② 回归插补。在单调缺失数据模式下，即变量$Y_1,\cdots,Y_{k-1}$全部有观测，Y_k对前r个有观测而丢失了后$n-r$个观测，用前r个完全观测个体计算Y_k对$Y_1,\cdots,Y_{k-1}$的回归，然后利用回归的预测值插补缺失值。回归替换法首先需要选择若干个预测缺失值的自变量，然后建立回归方程估计缺失值，即用缺失数据的条件期望值对缺失值进行替换。该方法利用了数据库中尽量多的信息，但也存在一些弊端：第一，容易忽视随机误差，低估标准差和其他未知性质的测量值，而且随着缺失信息的增多而变得更加严重；第二，研究者必须假设存在缺失值所在的变量与其他变量存在线性关系，很多时候这种关系是不存在的。③BUCK方法。将回归插补推广到更一般的缺失数据模式，首先基于完全观测的个体从样本均值和协方差阵估计均值和协方差，然后使用这些估计对每一种缺失数据模式计算缺失变量关于已观测到的变量的最小二乘线性回归，然后用回归预测值代替缺失值。在缺失机制是MCAR的假定下，如果分布矩满足一定条件，得到的是总体均值的相合估计；如果缺失机制是MAR的，在附加其他的一些假定

条件，估计仍然是相合的；但值得注意的是，如果用来预测缺失值的自变量值超出用于回归的自变量数据的范围，那么此时得到的缺失值的插补值将不可行。在 BUCK 方法下，同样会对总体的方差和协方差产生低估，但是比起无条件均值插补还是有所改善。

4. 热卡插补法

对于一个包含缺失值的变量，热平台插补是在已有数据中找到一个与它最相似的对象，然后用这个相似对象的值来进行填充。热卡插补中，不再假定缺失数据的预测分布，对有记录的数据，按照一定的概率抽取数值来插补缺失值。在等概抽样情况下，采用热卡插补的总体均值估计量可表示为：$\bar{y}_{HD}=\{r\bar{y}_R+(n-r)\bar{y}^*_{NR}\}/n$，其中 $\bar{y}_R$ 是响应单元的均值，$\bar{y}^*_{NR}$ 为未响应单元的均值；$\bar{y}^*_{NR}=\sum_{i=1}^{r}\frac{H_i y_i}{n-r}$，其中 H_i 是 y_i 用于代替缺失的 Y 值的次数，有 $\sum_{i=1}H_i=n-r$。从上述的表达式可以看出 $\bar{y}_{HD}$ 的性质依赖与产生的 $\{H_1,\cdots,H_r\}$ 所用的方法。因此，采用不同的方法从已记录单元中产生插补值就直接决定了 $\{H_1,\cdots,H_r\}$ 的分布是否存在，进一步决定了该方法下总体估计量的性质。

热平台插补是最流行的插补方法之一，因为简单直观，但不同的问题可能会选用不同的标准来对相似进行判定。与均值替换法相比，利用热卡填充法插补数据后，变量的标准差与插补前比较接近。在回归方程中，使用热卡填充法容易使回归方程的误差增大，参数估计变得不稳定，而且这种方法使用不便，比较耗时。

(1) 有放回简单随机抽样的热卡插补。在这种方法下，缺失值的插补值是通过对已记录数据进行有放回的简单随机抽样而获得。此时，$\{H_1,\cdots,H_r\}$ 的分布是一个多项分布，概率值为 $\{1/r,\cdots,1/r\}$，样本大小为 $n-r$。此时，$\{H_1,\cdots,H_r\}$ 的分布矩很容易求得，从而也可以直接得到总体均值和方差的估计值。假定缺失机制是 MCAR 的情况下，采用该种方法得到插补结果的均值是总体均值的无偏估计，但会造成方差的高估，而且高估的量不可忽略，尤其是当回答率为 50% 时，方差高估的量达到最大。为了改进被高估的方差，可以采用

一些例如简单随机抽样、限制对已回答数据的使用次数、对已回答数据进行排序并进行系统抽样等方法进行调整。

(2)分层热卡插补。不论是采用有放回还是无放回的简单随机抽样,所利用的信息仅仅是该变量的自身数据;为了提高估计效率,往往需要借助辅助信息。分层热卡插补就是像条件均值插补一样,按照某一个记录完全的变量对该变量进行分层,然后对分层后的数据进行热卡插补。

(3)最近距离热卡插补。根据测量单元间距离的函数,在临近缺失数据的有记录单元内,选择满足所设定距离条件的辅助变量对应的记录值作为插补值。特别地,分层热卡插补可以看作距离函数取值为 0、1 的最近距离热卡插补,其中 0 代表在同一层内,1 代表在不同层内,由于采用的函数比较复杂,对于其均值和方差性质的估计比较复杂,这里不做过多的讨论。

(4)序惯热卡插补:首先对数据进行分层,然后在每层中按照选定的某一个辅助变量排序,最后在其前后相邻的 10 个数据中,找到使得设定的某一个距离函数的值达到最小,用于构建距离函数的变量要求和将要插补的变量之间高度相关。更一般的情况下,可以采用其他的变量,但是要求距离函数的取值和该值所对应的缺失变量值被选作插补值的次数成正比。在获得插补值后,还要检查其是否满足一些基本的制约条件,避免犯一些逻辑上的错误。

5. 冷平台插补

冷平台插补与热平台插补方法不同,插入值是从以前的调查中或其他信息来源,譬如从历史数据中得到。有关冷平台插补方法的理论还较少,而且与前面插补方法的原理类似,也不能保证消除估计偏差。

(三)缺失数据处理的实际应用

1. 数据缺失机制的分析[①]。

收集 2006—2008 年全国除四个直辖市外的 282 个地级以上城

① 参考《SPSS 中的缺失值分析》, http://wenku.baidu.com/view/501692c7aa00b52acfc7cadc.html, 2009 年。

市的城市规模(市辖区人口数,记为 X)、人均 GDP(Y)数据,共 846 个观测值,X 和 Y 的相关系数为 0.3586,两者的散点图见图 4-1。

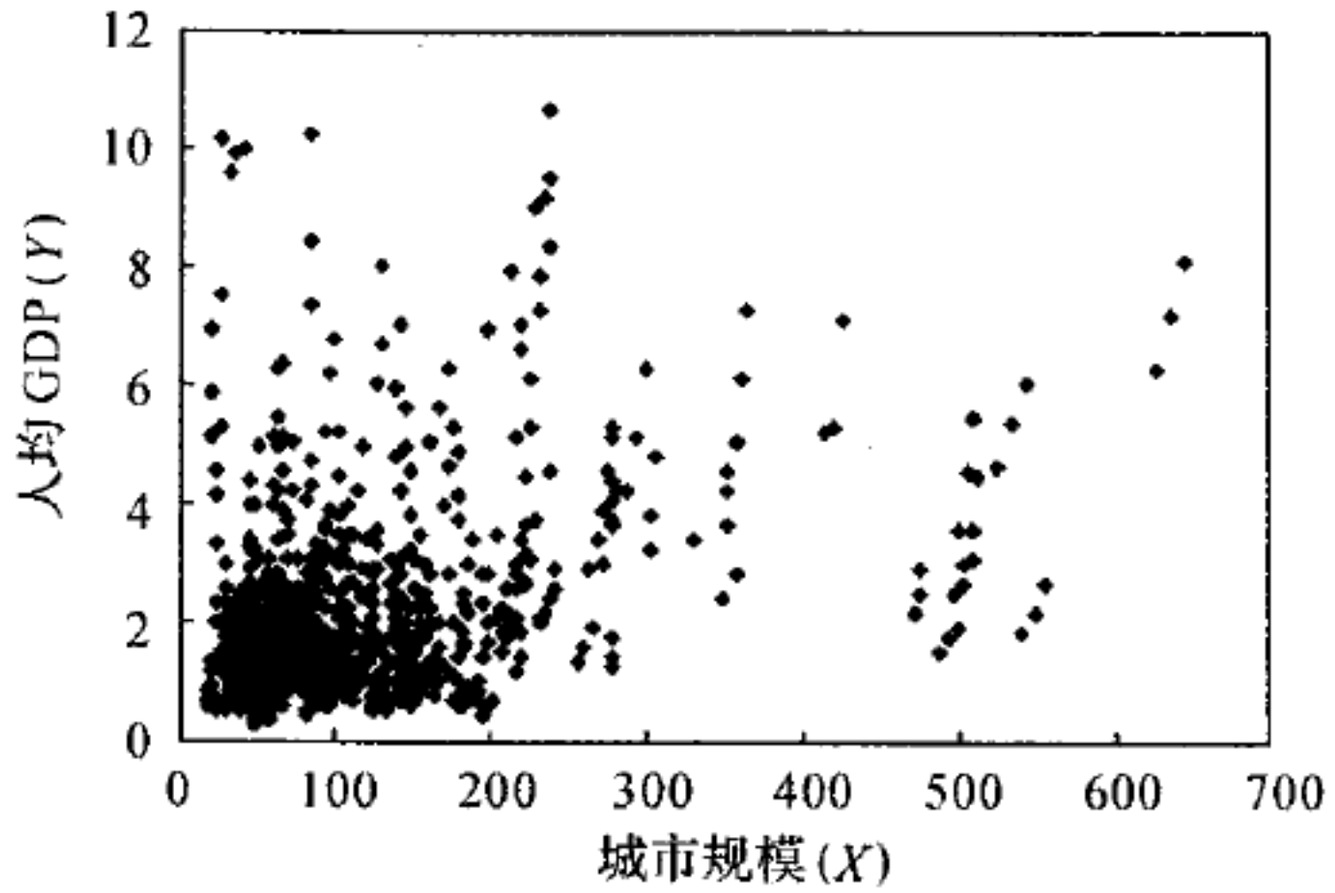

图 4-1　变量 X 与 Y 的散点图

(1)完全随机缺失(MCAR)。假设对 X 和 Y 随机地删除了大约 5%的观测值。在这种随机缺失机制下,期望结果数据为 MCAR。对存在缺失数据的结果变量 X_{miss} 和 Y_{miss} 进行相关分析,相关系数用基于列表删除的方法来计算,期望这个相关是已知相关系数 0.3586 的无偏估计。表 4-1 显示了 SPSS 相关分析的结果。

表 4-1　变量 X_{miss} 和 Y_{miss} 的相关系数

		X_{miss}	Y_{miss}
X_{miss}	Pearson Correlation	1.000	
	Sig. (2-Tailed)	0.000	
	N	804	
Y_{miss}	Pearson Correlation	0.3703**	1.000
	Sig. (2-Tailed)	0.000	0.000
	N	763	804

从表 4-1 可看到 X_{miss} 有 804 个有效观测值,Y_{miss} 有 804 个。经过列表删除以后,相关性用 763 个完整个案来计算。相关系数为 0.3703,与变量 X 与 Y 的已知相关系数 0.3586 较接近。

表 4-2 给出了 SPSS MVA 中用极大似然估计计算的相关系数,

以及 Little's MCAR 检验结果，p 值为 0.132，不能拒绝零假设，即数据缺失机制为 MCAR。

表 4-2　EM Correlations[a]

	X_{miss}	Y_{miss}
X_{miss}	1.000	
Y_{miss}	0.379	1.000

注：a. Little's MCAR test：Chisquare=4.054，df=2，Prob=0.132.

(2)随机缺失(MAR)。以下通过随机地设置城市规模(X)值大于 200 万人的个案中 40%的 Y 值共 42 个个案缺失，以判断 Y 值的缺失对 X 值是否是偶然的，形成缺失值结构为 *MAR*。

通过列表删除，Y_{miss} 和 X 之间的相关系数是 0.331(806 个个案)。图 4-2 显示出对 Y 变量缺失值进行均值置换的效果。

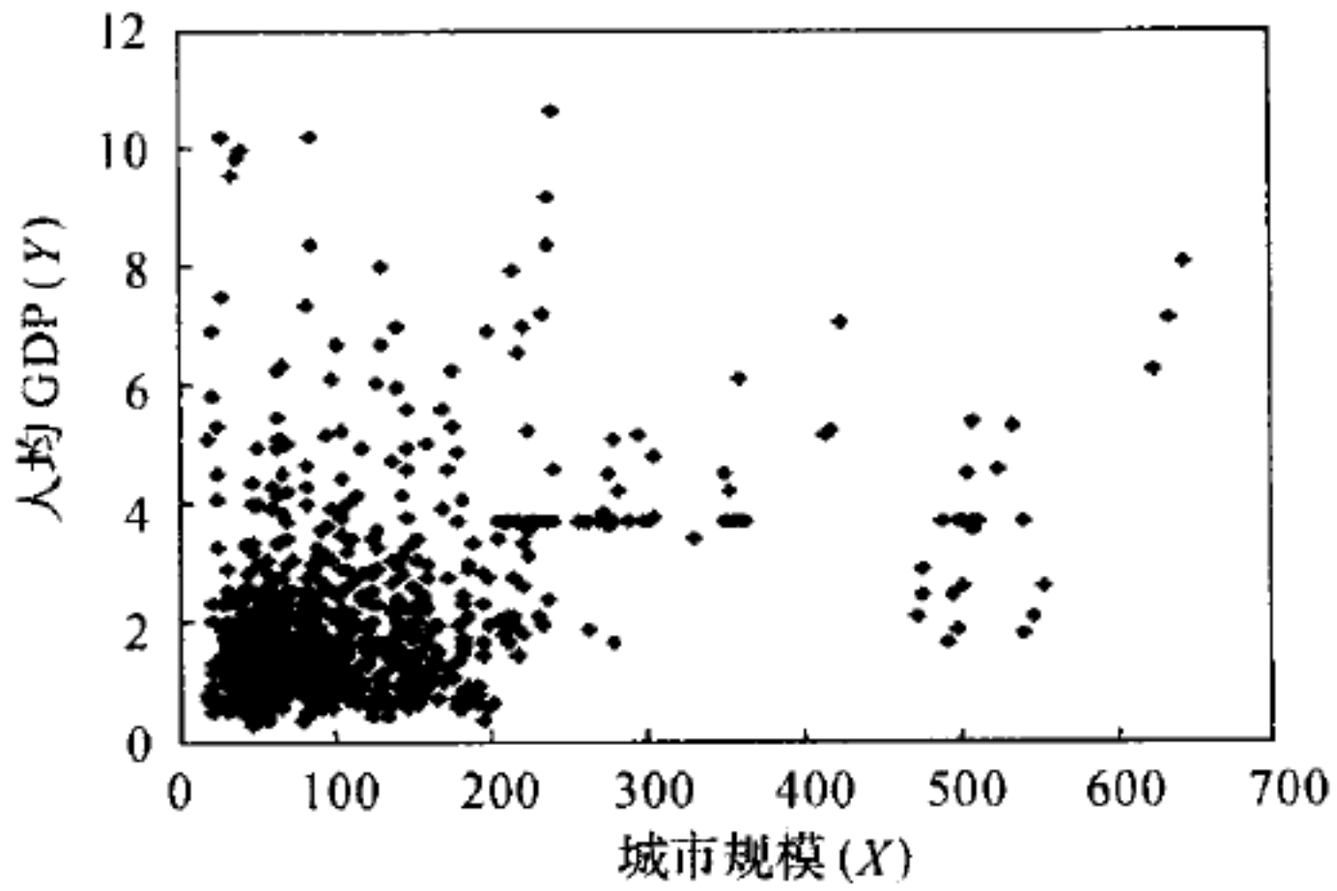

图 4-2　变量 *Y* 缺失值的均值替换效果

从图 4-2 可发现均值置换的一个主要问题：保持 Y_{miss} 的均值时，Y_{miss} 和 X 之间的相关性是曲解的，均值置换的 Y_{miss} 和 X 之间的相关系数改变为 0.3695(846 个观测量)。

如果不进行均值置换，而用 Y_{miss} 对 X 的回归构造一个回归方程，然后预测 Y_{miss} 缺失的个案的 Y_{miss}。图 4-3 显示了相应的替换结果。

当然，回归替换也存在一定的问题，一方面是回归估算值的方差

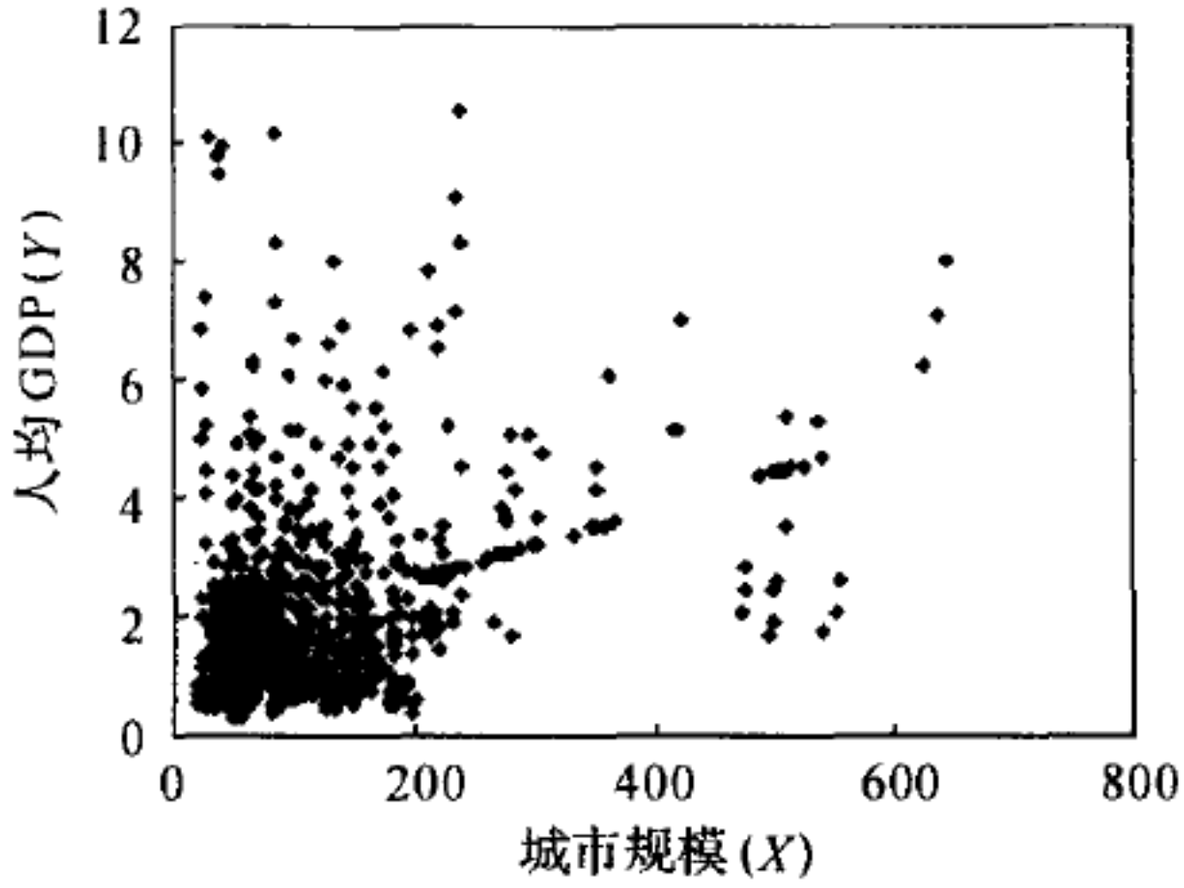

图 4-3　变量 Y 缺失值的回归替换效果

太小，另一方面回归估算的 Y_{miss} 和 X 之间的相关系数为 0.3962，夸大了 Y 和 X 的相关系数。

表 4-3　EM Correlations[a]

	Y_{miss}	X
Y_{miss}	1.000	
X	0.3576	1.000

注：a. Little's MCAR test：Chisquare=132.009，df=1，Prob=0.000.

表 4-3 给出了 SPSS MVA 中用极大似然估计计算的相关系数，估计值为 0.3576，与 Y 和 X 已知的相关系数 0.3586 非常接近。Little's MCAR 检验的卡方值为 132.009，p 值为 0.000，拒绝零假设，即数据缺失模式为 MAR。

2. 缺失数据处理的一些实际技巧

以上分析了缺失数据的处理方法、数据缺失机制的检验及应用，接着再给出对缺失值进行处理的一些实际技巧。

假设在分析一个商场销售数据时，发现有多个记录中的属性值为空，譬如顾客的收入(income)属性。对于这些为空的属性值，可以采用以下方法、技巧进行遗漏数据处理。

(1)忽略该条记录。若一条记录中有属性值被遗漏了，则将此条记录排除在数据分析过程之外，当然这种方法有时并不很有效，尤其

是当每个属性遗漏值的记录比例相差较大时。

(2)手工填补遗漏值。一般讲这种方法比较耗时,而且对于存在许多遗漏情况的大规模数据集而言,显然不具可行性。

(3)利用缺省值填补遗漏值。对属性的所有遗漏值均利用一个事先确定好的值来填补。例如,都用"OK"来填补。但当一个属性遗漏值较多时,若采用这种方法就可能误导统计分析结论。

(4)利用均值填补遗漏值。譬如计算收入(income)属性(值)的平均值,并用此值填补该属性所有遗漏的值。若顾客的平均收入(income)为 12000 元,则用此值填补 income 属性中所有被遗漏的值。

(5)利用同类别均值填补遗漏值。这种方法尤其适合能够对记录进行分类时加以使用,例如若要对商场顾客按信用风险进行分类分析时,就可以用在同一信用风险类别下(如良好)的 income 属性的平均值,来填补所有在同一信用风险类别下属性 income 的遗漏值。

(6)利用最可能的值填补遗漏值。可以利用回归分析、贝叶斯公式或决策树推断出该条记录特定属性的最大可能取值。例如利用数据集中其他顾客的属性值,可以构造一个决策树来预测属性 income 的遗漏值。

(7)推理插补。根据所得的信息推断缺失值,像先前调查的类似项目、目前调查中的相关项目等。例如,一个被调查者提供了三个孩子的名字,但"子女数"项空着,则可推出子女数为 3。

3. 缺失数据处理方法的比较

对于各种插补方法的应用,金勇进等(2009)在《缺失数据的统计处理》一书中做了很好的实证分析[①],在此不再重复开展类似的分析,而借用金勇进等(2009)的案例来加以说明。

(1)数据来源。分析数据取自中国质量管理协会和中国人民大学于 2003 年共同进行的"中国住宅产业住户满意度调查"。调查采

① 金勇进、邵军:《缺失数据的统计处理》,中国统计出版社 2009 年版,第 88—92 页。

用入户面访的方式，核心变量是“住宅总体满意程度”，用变量 Y 表示；剔除一些无回答单位，得到含有 5755 个“完整”单位的数据集，以下以此数据集为研究对象的总体。

为了进行有效插补，选取一些与变量 Y 有联系的变量：家庭平均月收入、居住户型、受教育程度、所在城市、居住小区、年龄等。

(2)数据整理。从 5755 个所谓“总体”的已知单元中，随机抽出 50%即 2878 个单元作为样本，按辅助信息“家庭平均月收入”和“居住户型”进行分组，构造插补层。其中将家庭平均月收入分为低、中、高 3 组，居住户型分为二室及以下和三室及以上 2 组，共形成 6 个交叉的插补组。

在 2878 个样本单元中，随机抽取 1783 个单元为回答单元，回答率约为 62%。不同的插补调整组中回答率有所差异，说明受访者是否回答与辅助变量收入水平、居住户型有一定的关系。

(3)插补结果与分析。以 5755 个单元的完全数据集为总体，从中抽取 2878 个单元为样本，在样本中有回答的单元数为 1783，其余为无回答单元。采用 6 种插补方法对无回答单元进行插补，形成插补以后的 2878 个样本单元“完全数据集”，并对插补结果进行分析与评论。6 种插补方法分别是：单一均值插补、分层均值插补、随机热卡插补、序贯热卡插补、回归插补①、回归插补②。各种方法的说明及插补结果如表 4-4 所示。

从表 4-4 中的插补结果可以发现：(1)在插补法中，利用辅助信息对样本单位进行分层有利于改善估计量性质，减少无回答误差；(2)由于利用了更多的辅助信息，序贯热卡插补的效果通常好于随机热卡插补；(3)如果找到正确的辅助变量，回归插补可以收到较好的估计效果；(4)在回归插补中，遗漏重要辅助信息会降低插补的效果。

表 4-4　几种插补方法的结果比较

方　法	N(N)	MEAN	DEVIATION	STD. V	方法说明
总体均值 $\bar{Y}$	5755	6.9567	0	1.7207	直接用 5755 个观测值得到的结果
单一均值插补（$\bar{Y}_R$）	2878	7.0051	0.0483	1.3868	利用回答单元均值插补所有缺失值
分层均值插补（$\bar{Y}_{mean}$）	2878	6.9494	−0.0074	1.4059	利用收入和户型变量将样本单位划分为6层，用各层回答单位的均值插补该层内所有缺失值
随机热卡插补（$\bar{Y}_{ran}$）	2878	6.9846	0.0278	1.8012	利用收入和户型变量将样本单位划分为6层，在每一层内随机选取回答单位的均值对同层缺失值进行插补
序贯热卡插补（$\bar{Y}_{sq}$）	2878	6.9620	0.0053	1.7867	利用收入和户型变量将样本单位划分为6层，并在每一层内按照居住小区变量进行排序。对每个缺失值都利用其前一个回答值进行插补
回归插补①（$\bar{Y}_{reg1}$）	2875	6.9470	−0.0098	1.4190	选取与变量 Y 相关性最高的收入、户型、教育程度和所在城市四个变量作为自变量建立回归方程，以回答数据进行拟合。利用回归预测值对无回答值进行插补
回归插补①（$\bar{Y}_{reg2}$）	2875	6.9849	0.0282	1.4093	剔除收入，用其他三个变量作为自变量建立回归方程，以回答数据进行拟合。利用回归预测值对无回答值进行插补

资料来源：转引自金勇进、邵军：《缺失数据的统计处理》，中国统计出版社 2009 年，第 91 页。