

Multiple Imputation:
Issues and
Guidance for Practice

缺失数据的 多重插补

应用案例与软件操作

严洁 / 著



重庆大学出版社

<http://www.cqup.com.cn>

万卷方法 / 社会科学研究方法 · 前沿与应用

缺失数据的多重插补

——应用案例与软件操作

严 洁 著

重庆大学出版社

图书在版编目 (CIP) 数据

缺失数据的多重插补: 应用案例与软件操作 / 严洁
著. —重庆: 重庆大学出版社, 2016.4
(万卷方法)
ISBN 978-7-5624-9635-9

I. ①缺… II. ①严… III. ①社会科学—定量分析—
数据处理—研究 IV. ①C3-39

中国版本图书馆CIP数据核字 (2016) 第034130号

缺失数据的多重插补

——应用案例与软件操作

严洁 著

策划编辑: 林佳木

责任编辑: 文鹏 邬小梅 林佳木 版式设计: 林佳木
责任校对: 关德强 责任印制: 赵晟

*

重庆大学出版社出版发行

出版人: 易树平

社址: 重庆市沙坪坝区大学城西路21号

邮编: 401331

电话: (023) 88617190 88617185 (中小学)

传真: (023) 88617186 88617166

网址: <http://www.cqup.com.cn>

邮箱: fxk@cqup.com.cn (营销中心)

全国新华书店经销

重庆市国丰印务有限责任公司印刷

*

开本: 940mm × 1360mm 1/32 印张: 9.375 字数: 185千

2017年2月第1版 2017年2月第1次印刷

印数: 1—3 000

ISBN 978-7-5624-9635-9 定价: 38.00 元

本书如有印刷、装订等质量问题, 本社负责调换

版权所有, 请勿擅自翻印和用本书
制作各类出版物及配套用书, 违者必究

社会科学研究方法·前沿与应用

编 委 会

学术委员会（按汉语拼音为序）

风笑天 连大祥 刘涛雄 卢春龙 沈明明

汤维强 唐世平 王天夫 张小劲 朱云汉

编委会（按汉语拼音为序）

陈 硕 洪 伟 呼和那日松 吕 杰

孟天广 庞 珣 任莉颖 苏毓淞 孙 龙

严 洁 张小劲 郑 路 周 强

总 序

中国当代社会科学各学科的研究方法多为西方舶来，这一事实不可逃避。由此亦衍生出不尽如人意的另一个事实，即这些滥觞于西方的科学研究方法在国内学界的传播与应用，并未能像各个学科的一般理论和知识那样，为国内学者或学生广泛习得并恰当使用。归根究底，造成这种现状的原因无外乎如下两点：一是中国社会科学诸学科长期以来重描述的研究范式；二是青年研究者在初入社科研究领域时（多为文科生）就已产生了视数学为畏途的心态。以此为背景的研究或秉承旧有的研究范式，对新的研究方法视而不见，或对西方学界主流的研究方法进行囫圇吞枣、生吞活剥式的使用。这都不利于正在迅速发展的中国社会科学与国际学术界交流。这个世纪开始，随着大批海外学子陆续回国任教，他们给社会科学学科发展带来很大冲击，也从方法上改变了学科研究的范式。

这里说的方法变革所造成的范式转变，并非简单的定量与定性方法之争。此处所谓新的范式是指通过严谨的研究设计，应用科学的方法（定量或定性方法皆属之），采集数据去从事研究。过去数十年国内学界所累积的学科知识，在这个新的研究范式下，已得到验证，甚至有突破性的理论创新。不过，任何转变都不可

能一蹴而就，固有研究范式的惯性及其导致的对于科学方法的疏离感，仍然如紧箍咒一般，时刻影响着科学研究方法在国内社会科学学科的发展。

“社会科学研究方法·前沿与应用”丛书的主要目的就是希望拉近学生与科学方法的距离，从而激发他们学习并应用这些方法。2012年暑期，重庆大学出版社的雷少波编辑做客清华大学政治学系，言谈间，我提出组编一套介绍社会科学研究方法丛书的想法，有别于坊间同类书籍介绍方法的丛书，这套丛书的最大特色在于将理论背后的“理论”呈现给读者，让初学者知其然，更能够知其所以然；同时使用实际的研究成果，提供最直观的应用案例，让读者了解到这些研究方法如何应用到以中国为主题的研究，通过主题的熟悉感，真正拉近读者与方法之间的距离，提升读者学习的热情，为有志于从事相关领域研究的人士提供扎实、恳切的引导及参考。

这套丛书编写的体例大致分为理论介绍、案例点评和软件实操展示三个部分。第一部分的理论介绍，主要是编写者结合自身学习该方法的历程，用最浅显易懂的方式（尽可能使用最少或最基本的数学），来介绍方法背后的理论与原理，以及应用该方法的场景，并重点介绍该方法的标准操作流程，以及提示该方法在操作上可能遇到的限制，为读者学习理论提供最大的便利。第二部分的案例点评是本书的亮点，编写者选取应用该方法的范文（优先选择以中文撰写和使用中国数据的论文）进行点评，这样选择的目的是如前所述，是希望读者通过既有的研究，切实学习到

该方法是如何在中国场景下落地的。编写者的点评首先发挥导读的作用，同时亦不会避重就轻，而是无褒贬之忌讳，既点出该方法在范文中使用的准确、巧妙之处，也会点出研究过程中存在的短板，不论是方法操作上的问题，还是研究设计本身的缺陷，抑或是数据上造成的罅隙，这也是编写者希望读者能够在实操过程中尽量避免的部分，毕竟方法的习得容易，但是要用得对、用得好，还是需要熟悉方法、经验丰富的资深使用者指导。此部分的点评就是希望用文字为读者建构一个面对面的学习场景，起到教师与学生一对一现场教学的效果。第三部分的软件实操展示，使用的软件主要还是坊间常用的 SPSS、STATA、R 三个软件。编写者使用数据范本，实际演示该方法通过不同软件的操作步骤，尽可能做到精准、简洁，同时兼顾可读性、可视性。原则上，这部分会至少涵盖两个软件操作展示，未来如读者有需求，我们将在丛书的网页以附录的方式，编写该方法在其他软件的操作步骤。

丛书总序的最后，必须感谢清华大学的张小劲老师、王天夫老师、洪伟老师、孟天广老师，北京大学的严洁老师、任莉颖老师，中国政法大学的卢春龙老师，人民大学的孙龙老师和远赴英国的呼和那日松老师，诸位同仁鼎力相助，为这套丛书的出炉提供了智力和时间上的大力支持。丛书编著过程中，经历了许多波折与挑战，若没有他们的投入，这套丛书的顺利问世几乎是不可想象的。特别感谢重庆大学出版社雷少波编辑的鼎力支持，我当初言谈间大胆的设想方能实现，也特别感谢林佳木编辑、邹荣编辑不厌其烦地催稿，如果说雷编辑是伯乐的话，林、邹两位编

辑则是催促着我们前进的鞭子，丛书的完成，这些文字背后的贵人缺一不可。

编著这套丛书，是为试图改进中国社会科学研究对于国际主流方法的使用状况，这也是上述诸位共谋此事背后的雄心壮志，但不积跬步无以至千里，任何实质性的转变都要依靠些微的积累和推进方能实现，落在实处，就是这一本本小书的质量。若能对有需要的读者有所启发，或能由此催生后续研究的跟进，就是编写者最大的心愿。由于编写者学识所限，这套丛书一定还存在诸多不足和有待完善之处，希望读者和相关方法的使用者及时反馈批评意见，以便于我们再版时加以订正和改进。

苏毓淞

于清华园

2016.10.5

序

我的同事严洁博士请我为她的一本关于研究方法的新书作序，我自然欣然从命。

这本书是重庆大学出版社组织的万卷方法系列丛书中的一本。看到这套丛书面世，我欣喜之余，又不禁有些感慨。犹记得二十年前，我曾经试图将美国世哲出版公司（Sage Publications）的那一套著名的研究方法丛书翻译介绍到中国来，但没有成功——彼时并没有出版社愿意主动出版，若坚持出版则要我支付出版社高额费用。

同样是二十年前，我曾在北大接待时任美国政治学会（American Political Science Association）的主席 M. Kent Jennings 教授（他是我在密西根大学的导师之一）的访问。临走时他对我说了一句话：“你在这里没有同事，我担心你学术上的进步会很困难……”

那时“海归”还不多，获得“同事”的途径只有一个——自己培养。因此我在北京大学中国国情研究中心开展与密西根大学社会研究院（Institute for Social Research, University of Michigan）合作，开设了一系列暑期培训项目，连续四年在北大和密西根为国内大学和研究机构的研究生和青年教师提供实证和

量化研究方法方面的培训——严洁博士就是最早的一批学员之一，而且是其中的佼佼者。

经过这么多年的学习和实践，她已经成长为一个在国际国内学界得到普遍认可的研究方法方面的专家。诚然，相当多在国外接受教育的博士都受过研究方法方面的系统培训，但他们普遍缺乏实践经验。严洁的出众之处在于，她既在国内国外接受过系统的训练，又在二十年来参与了北大中国国情研究中心几乎所有的实证和量化分析研究项目，积累了丰富的实践经验，具备了独立主持大型科研项目的素质和能力，特别是在数据管理和数据分析方面颇有造诣。

严洁博士这本关于多重插补（Multiple Imputation）的书，是她这些年在数据管理和数据分析方面经验的总结，某种意义上说，也是她博士论文研究的延续。我很欣慰看到一个青年学者在学术上日渐走向成熟，并且在方法上逐步进入非常细致的领域。

多重插补是个非常繁复且细致的工作，它不单单是数据处理的一种技术方法，更要求联系研究项目本身的方方面面加以注意。而严洁这本书有两个突出的特点，为读者深入系统地了解 and 掌握它提供了可能。第一，她为缺乏实践经验的学生和读者提供了大量而具体的实例，使他们得以理解这种方法的意义并学会如何运用；第二，她详细地介绍了运用现有计算机软件完成这项工作的方法，使琐细繁复的工作变得易于操作，从而能够在教学 and 实际工作中给学生和研究者提供针对无回答问题的解决办法和很

好用的工具。我希望，也相信，这本书能够为更多学生和研究者所用，成为案头工具书。

最后，也希望万卷方法这套丛书能够成功，并且可以继续拓展开去。

是为序。

沈明明

燕园

2016年3月2日

作者自序

近些年来，越来越多的社会科学领域学者应用社会调查的方法来探讨中国政治、经济、社会发展过程中的理论和实践问题。但是在社会调查的数据采集过程中，由于各种原因而无法获得某个样本的任何一项回答或者无法获得样本对某个或某些个问题的回答的情况越来越多，前者称为单元无回答/单元无应答，后者称为项目无回答/题目无应答。由无回答引起的数据缺失对于后期的数据分析造成很大的影响，为此，在使用统计分析方法进行描述和推断之前，需要考虑缺失值的处理问题。

对于缺失值的处理有多种方法，其中多重插补(multiple imputation)是近些年来最广为接受的一种。简单地说，多重插补就是给每个缺失单元插补上多个值。主要包括三大步骤，即插补、分析、综合。那么接下来，我们会产生一系列疑问：

1. 一个缺失的单元为什么可以给出多个插补值？
2. 这些插补值是怎么计算来的？
3. 既然要插补多个值，那么需要插补多少个比较合适？
4. 插补后的数据集该怎么用？

5. 什么情况下可以用多重插补?

6. 怎样借助统计软件来作多重插补?

本书的主要目的就在于借助实例来解答这些问题。本书包括三个部分,在第一部分主要介绍多重插补的基本原理和方法、步骤;第二部分则借助四篇精心挑选出来的文章,通过对文章的评析来进一步解答有关多重插补方法的细节问题和相应的注意事项;第三部分则讲解和示例如何用 SPSS、STATA 这两种统计软件来实现多重插补。

本书适用于从事定量研究的学者和学习定量研究的硕博研究生。虽然书中的例题属于社会科学领域,但是对于公共卫生领域、环境科学领域的抽样调查数据的缺失值处理也具有一定的参考价值。

本书在写作的过程中得到了很多老师和朋友的帮助。感谢我的博士生导师、北京大学政府管理学院的沈明明教授和杨明教授。两位教授从 1995 年开始带着我进行社会调查,引导我进入了政治学的定量研究领域,在北京大学中国国情研究中心的所有调查项目中都给予了我学习和成长的机会,花费大量精力培育了我在这个学科领域内的学识和视野。感谢北京大学社会学系的邱泽奇教授,他于 2008 年邀请我参与北京大学中国社会科学调查中心的调查工作,使得我较早参与到全国大规模的追踪调查项目中,邱教授在本书的写作过程中也提出了诸多宝贵的意见和建议。感谢北京大学医学部的黄悦勤教授,她的引导和建议使得我的研

究能够向公共卫生领域内拓展。感谢我的同窗好友、北京大学中国社会科学调查中心的任莉颖副研究员，她一直在我诸多的学术研究中给予大量的支持和帮助。感谢清华大学政治学系的张小劲教授、苏毓淞副教授，两位教授的不断鼓励给予了我信心，使得我得以按时完成书稿。感谢清华大学政治学系的孟天广副教授、北京大学政府管理学院的黎娟娟博士，两位对本书提出了非常具体和细致的建议。感谢在匿名评审过程中给予本书批评和建议的老师们，这些意见使得本书更加完善。最后要感谢重庆大学出版社万卷方法丛书的雷少波和林佳木编辑，这本书的体例策划得益于两位编辑的卓见，林佳木、邬小梅两位编辑在本书的成稿过程中对文字的校改付出了大量的劳动，在此表示衷心感谢。

希望本书能对读者们的量化研究有所裨益，欢迎读者们批评指正。

严 洁

2016年2月于北京大学

目 录

导 言 社会调查、无回答与缺失数据	/ 1
第 1 章 删除法和单一插补法	/ 17
1.1 删除法	/ 18
1.1.1 列表删除	/ 18
1.1.2 成对删除	/ 19
1.2 单一插补法	/ 22
1.2.1 均值插补	/ 22
1.2.2 回归插补	/ 25
1.2.3 热平台插补	/ 28
1.2.4 冷平台插补	/ 34
1.2.5 LOCF 与 BOCF 法	/ 34
第 2 章 多重插补法概述	/ 37
2.1 多重插补的基本原理	/ 38
2.1.1 什么是多重插补	/ 38
2.1.2 为什么一个缺失单元可以有多个插补值	/ 42
2.1.3 缺失类型	/ 44
2.1.4 多重插补的方法	/ 53
2.1.5 进行多少次插补才有效	/ 64
2.1.6 多个插补值怎样使用	/ 67
2.1.7 什么情况下用多重插补	/ 69
2.2 多重插补的发展简史	/ 70
2.2.1 起始阶段 (1977—1987 年)	/ 71
2.2.2 第二阶段 (1988—1997 年)	/ 73
2.2.3 第三阶段 (1998 年至今)	/ 77

第 3 章 多重插补的方法、步骤	/ 81
3.1 准备插补变量	/ 82
3.1.1 初步设定分析模型	/ 82
3.1.2 初选插补模型的变量	/ 86
3.2 检验和确定插补模型	/ 91
3.2.1 选择插补方法	/ 91
3.2.2 变量转换	/ 92
3.2.3 多重共线性诊断	/ 99
3.2.4 确定插补次数	/ 101
3.2.5 执行多重插补, 并诊断插补模型	/ 106
3.3 多重插补后的统计分析	/ 107
3.3.1 插补后进行简单随机抽样下的统计分析	/ 108
3.3.2 插补后进行复杂抽样下的统计分析	/ 111
3.4 小结	/ 116
第 4 章 应用案例分析	/ 119
4.1 政治学应用案例	/ 121
4.2 教育学应用案例	/ 135
4.3 经济学应用案例	/ 172
4.4 管理学应用案例	/ 195
第 5 章 使用 SPSS、STATA 软件做多重插补	/ 239
5.1 使用 SPSS 软件做多重插补	/ 240
5.2 使用 STATA 软件做多重插补	/ 253
5.2.1 多重插补前的准备工作	/ 253
5.2.2 多重插补	/ 257
5.2.3 多重插补后的统计分析	/ 263
5.3 SPSS、STATA、SAS、R 软件做多重插补的异同	/ 271
参考文献	/ 274

／ 导言 ／

社会调查、无回答与缺失数据