

外国优秀统计学教材系列丛书 · 翻译版



作者

[美] Rodenick J.A. Little
& Donald B. Rubin

翻译

孙山泽

校审

耿直

全国统计教材编审委员会组织引进

缺失数据统计分析

Statistical Analysis With Missing Data



中国统计出版社
China Statistics Press

<http://www.stats.gov.cn>

外国优秀统计学教材系列丛书， 翻译版

全国统计教材编审委员会组织引进

缺失数据统计分析

Statistical Analysis With Missing Data

(第二版)

[美] Roderick J.A.Little
& Donald B. Rubin 著

孙 山 泽 译
耿 直 校审

 中国统计出版社
China Statistics Press

Statistical Analysis With Missing Data (2nd Ed.)

Roderick J.A. Little & Donald B. Rubin

Copyright ©2002 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

ISBN 0-471-18386-5

All rights Reserved. Authorized translation from the English language edition published by John Wiley & Sons, Inc.

本书中文简体字版由 John Wiley & Sons, Inc. 授权中国统计出版社独家出版。未经出版者书面许可，不得以任何方式翻印、拷贝、仿制或转载。

版权所有，侵权必究。

北京市版权局著作权合同登记号：图字：01-2003-3079

(京)新登字 041 号

图书在版编目(CIP)数据

缺失数据统计分析

/(美)利特尔(Little R.J.A.), (美)鲁宾(Rubin, D.B.) 著. - 孙山泽 译

北京: 中国统计出版社, 2004.9

ISBN 7-5037-4495-2

I. 缺… II. ①利… ②鲁… ③孙… III. 统计数据-统计分析(数学)-教材 IV. C811

中国版本图书馆CIP数据核字(2004)第0883824号

责任编辑 / 杨映露 郭栋

封面设计 / 刘国宁

E-mail / cbsebj@stats.gov.cn

出版发行 / 中国统计出版社

通信地址 / 北京市西城区月坛南街75号

邮政编码 / 100826

办公地址 / 北京市丰台区西三环南路16号

电话 / (010) 63459084 63456600-22500(发行部)

印刷 / 北京兴伟印刷厂

经销 / 新华书店

开本 / 850 × 1168 毫米 1/16

字数 / 370 千字

印张 / 21

印数 / 1-3000 册

版别 / 2004 年 9 月第 1 版

版次 / 2004 年 9 月北京第 1 次印刷

书号 / ISBN 7-5037-4495-2/O.52

定价 / 38.00 元

中国统计版图书，如有印装错误，本社发行部负责调换。

引进外国优秀统计学教材专家委员会

主任：贺铿 国家统计局副局长

副主任：王吉利 国家统计局统计教育中心主任

谢鸿光 国家统计局办公室主任

严建辉 中国统计出版社副社长

委员：（按姓氏笔画排序）

文兼武 国家统计局统计科学研究所 所长
孙山泽 北京大学数学科学学院概率统计系 教授
朱维盛 中国统计出版社 副总编
吴喜之 中国人民大学统计学系 教授
何书元 北京大学数学科学学院概率统计系 教授
张尧庭 上海财经大学经济学院 教授
张润楚 南开大学数学学院统计学系 教授
邱东 东北财经大学 教授
陈江 美国雪城大学管理学院 教授
陈家鼎 北京大学数学科学学院概率统计系 教授
杨振海 北京工业大学应用数理学院 教授
孟晓犁 哈佛大学统计系 教授
林正炎 浙江大学理学院数学系 教授
茆诗松 华东师范大学统计系 教授
郑明 复旦大学管理学院统计系 副教授
柯惠新 北京广播学院 教授
徐国祥 上海财经大学统计学系 教授
袁卫 中国人民大学 教授
蒋萍 东北财经大学统计系 教授
谢邦昌 台湾辅仁大学统计资讯学系 教授
濮晓龙 华东师范大学统计系 副教授

办公室：

温明 国家统计局统计教育中心教材处 处长
陈悟朝 中国统计出版社第二书籍编辑部副主任
孙洪娟 国家统计局统计教育中心教材处 统计师
杨映霜 中国统计出版社第二书籍编辑部 编审

73248/01

出版说明

21 世纪的竞争是人才的竞争，是全球性人才培养机制的较量。如何培养面向现代化、面向世界、面向未来的高素质的人才成为我国人才培养的当务之急。为此，教育部发出通知，倡导在全国普通高等学校中使用原版外国教材，进行双语教学，培养适应经济全球化的人才。

为了响应教育部的号召，促进统计教材的改革，培养既懂统计专业知识又具备较高英语语言能力的统计人才，全国统计教材编审委员会在国家统计局领导的大力支持下，组织引进了这套“外国优秀统计学教材”。

为了做好“外国优秀统计学教材”引进工作，全国统计教材编审委员会将其列入了“十五”规划，并成立了由海内外统计学家组成的专家委员会。在对国外统计学教材的使用情况进行了充分了解，对国内高等院校使用外国统计学教材的需求情况进行了仔细分析，并对从各种渠道推荐来的统计教材进行了认真审定的基础上，制定了引进教材书目。在确定引进教材书目的过程中，我们得到了国内外有关专家、有关院校和外国出版公司及其北京办事处的支持和帮助，在此致谢。中国人民大学统计学系的吴喜之教授不仅推荐了大量的优秀候选书目，而且校译了影印教材的翻译目录，为这套教材的及早出版作了大量的工作，我们表示衷心的感谢。

这套引进教材多数是国外再版多次、反响良好，又比较适合国内情况、易于教学的统计教材。我们希望这套引进教材的出版对促进我国统计教材的改革和高校统计学专业双语教学的发展能够起到重要的推动作用。

全国统计教材编审委员会

2002年8月28日

前 言

早

在十九世纪 70 年代,具有缺失值数据的统计分析已相当兴旺,藉着计算机技术的优势得到迅猛发展,计算机技术已使从前吃力的数值计算成为一个简单的事情。这本书的目的是为了了解现在处理缺失数据问题的方法技术,为有缺失数据的分析给出基于似然的理论,这样可以系统化这些方法,并为未来的发展提供基础。本书的第一部分讨论历史上在三个重要的统计领域中缺失值问题的方法。这三个领域包括:有计划试验的方差分析,抽样调查和多元分析。这些方法虽然不是没有价值,但倾向于有一个就事论事的特征,常常由实践者给出解答,对理论性质的研究有限。第二部分对具有缺失值的数据的分析给出系统的方法。推断是基于似然函数,似然是根据数据构成及缺失数据机制的形式化统计模型求得的。第三部分给出了这些方法在各方面的应用,涉及回归、因子分析、列联表分析、时间序列和抽样调查等。第一部分中的许多历史上的方法能够由这些基于似然的方法导出(或近似),可作为一些例子。

本书是面对应用统计学家的,因此更强调例子,而不强调正则条件的精确叙述或定理的证明,但仍希望读者熟悉基于似然的推断的基本原则,在 6.1 节有点简短的回顾。这本书也假定读者了解完全数据分析的标准模型——正态线性模型,计数数据的多项模型,以及标准统计分布的性质,特别是多元正态分布的性质。某些章节还假定对统计活动的一些特殊领域是熟悉的,如试验设计的方差分析(第 2 章),抽样调查(第 3、4 和 5 章),列联表的对数线性模型(第 13 章)。一些说明性的例子也会引出其他的统计论题,如因子分析或时间序列(第 11 章)。这些例子的讨论是自我包容的,不要求很详细的知识,当然这些知识会增强读者对主要的统计论题的鉴赏力。我们曾经在一个 40 小时的统计课程中,安排了大约本书的四分之三的材料。

本书的第一版写于十九世纪 80 年代中期,其缺点是缺失数据方法主要由参数点估计和近似标准差以及大样本理论下的区间估计和检验推导构成。从那时

2 前 言

起,模拟后验分布的贝叶斯方法已得到广泛的发展,在第二版中反映了这些发展。与多重借补紧密相关的技术也比第一版得到更多的强调,更加重视它在处理缺失数据的理论和实践以及商业软件中的作用。本书的第一部分做了重组,增进了材料的来源。第二部分包含了 EM 算法的一些推广,它们在第一版时尚不可能利用,也包含了更多的贝叶斯理论和计算,在许多统计领域它们已成为标准的工具。似然方法的应用集中在新的第三部分中,数据不完全时,模型假定的诊断检验工作仍相当粗略。

由于第二版有一些主要的增加和修订,这儿提供一个表,说明原来第一版中的材料所处的位置。

第一版

1. 概念
2. 试验中的缺失数据
- 3.2. 完全个体完全分析
- 3.3. 可用个体分析
- 3.4. 填充缺失值
- 4.2.,4.3. 具有和不具有缺失值的随机化推断
- 4.4. 加权方法
- 4.5. 借补程序
- 4.6. 有不响应时抽样方差的估计
5. 基于似然函数的推断理论
6. 因子化似然方法
7. 一般模式缺失数据的极大似然
8. 正态例子的 ML
9. 未完全分类列联表
- 10.2. 一般的位置模型
- 10.3.,10.4. 推广
- 10.5. 稳健估计
11. 不可忽略模型
- 12.1.,12.2. 调查的不响应
- 12.3. 可忽略不响应模型
- 12.4. 多重借补
- 12.5. 不可忽略的不响应

第二版

1. 概念
2. 试验中的缺失数据
- 3.2. 完全个体分析
- 3.4. 可用个体分析
- 4.2. 从预测分布借补均值
已删去
- 3.3. 加权的完全个体分析
4. 借补
5. 借补不确定性的估计
6. 基于似然函数的推断理论
7. 因子化似然方法,忽略缺失数据机制
8. 一般模式的极大似然
- 9.1. 基于信息阵的标准误差
11. 多元正态的例子
13. 未完全分类列联表
14. 有缺失值的正态和非正态混合数据
12. 稳健估计
15. 不可忽略模型
- 3.3. 加权的完全个体分析
4. 借补
- 5.4. 多重借补的介绍
10. 贝叶斯和多重借补
- 15.5. 正态模式混合模型

从第一版以后,缺失数据方面的统计文献无论在应用上还是方法的发展都有了极大的扩充,这使得我们不可能调查全部的统计工作,另外也要使本书有一个可容忍的长度。我们倾向于在自己的实验范围内讨论应用,并集中注意力于贝叶斯方法和基于似然的方法,我们相信这会为应用提供一个强有力的理论基础,像基于广义估计方程之类的其他方法则留给另外的著作了。

在完成本书的过程中,许多人给予了帮助,要向他们致射: NSF 和 NIMH(通过许可 NSF—SES—83—11428, NSF—SES—84—11804, NIMH—MH—37188, DMS—9803720 和 NSF—0106914)帮助支持了本书研究报告的一些背景。第一版时, Mark Schluchter 帮助进行计算, Leisa Weld 和 T. E. Raghunathan 细心阅读了最后的文稿并提出一些很有帮助的建议, UCLA 课题 Biomathematics M232 以及 Harvard 大学课题 Statistics 220r 中我们的学生也提供了有益的建议。Judy Siesen 打印了我的许多草稿。Bea Shube 给予了热情的支持与鼓励。在第二版,要特别感谢 Chuanhai Liu 帮助进行计算, Mingyao Li, Fang Liu 和 Ying Yuan 帮助举例。许多读者帮助发现了一些打印及其它的错误,这中间要特别感谢 Adi Andrei, Samantha Cook, Shane Jensen, Elizabeth Stuart 和 Daohai Yu。

在完书后,我们继续发现许多统计问题看成缺失值问题是很有益的,即使数据集有完整的记录,而且缺失数据的研究可以是学习一般统计的一个很好的跳板。我们希望我们的读者能同意我们的观点并从本书获得激励。

R. J. A. Little

D. B. Rubin

目 录

前言

第一部分 回顾与基本方法

第一章 概论 3

- 1.1 缺失数据的问题 3
- 1.2 缺失数据模式 3
- 1.3 导致缺失数据的机制 10
- 1.4 缺失数据方法的分类 16

第二章 试验中的缺失数据 20

- 2.1 引言 20
- 2.2 完全数据的精确的最小二乘解 21
- 2.3 缺失数据的正确的最小二乘分析 22
- 2.4 填充最小二乘估计 23
 - 2.4.1 耶兹(Yates)方法 23
 - 2.4.2 使用公式计算缺失值 24
 - 2.4.3 迭代寻找缺失值 24
 - 2.4.4 用缺失值协变量的协差分析(ANCOVA) 25
- 2.5 Bartlett 的 ANCOVA 方法 25
 - 2.5.1 Bartlett 方法的有效性质 25
 - 2.5.2 记号 25
 - 2.5.3 参数和缺失的 Y 值的 ANCOVA 估计 26
 - 2.5.4 残差平方和和 $\hat{\beta}$ 的协差阵的 ANCOVA 估计 26
- 2.6 仅使用完全数据方法由 ANCOVA 获得缺失值的最小二乘估计 27

2.7 标准差的正确的最小二乘估计和一个自由度的平方和 30

2.8 多于一个自由度的最小二乘平方和 31

问题 33

第三章 完全个体和可用个体的分析,包括加权方法 35

3.1 引言 35

3.2 完全个体分析 35

3.3 加权的完全个体分析 37

3.3.1 加权调整 37

3.3.2 由不响应加权增加的方差 43

3.3.3 后分层和向已知边缘的校准 43

3.3.4 加权数据的推理 45

3.3.5 加权方法的总结 46

3.4 可用个体分析 46

问题 48

第四章 单一借补方法 51

4.1 引言 51

4.2 从预测分布借补均值 53

4.2.1 无条件均值借补 53

4.2.2 条件均值借补 53

4.3 从预测分布中抽取借补值 55

4.3.1 基于明确模型的抽取 55

4.3.2 基于模糊模型的抽取 57

4.4 结论 62

问题 62

第五章 借补不确定性的估计 65

5.1 引言 65

5.2 由单一填充数据集提供有效的标准误差的借补方法 66

5.3 用再抽样的借补数据的标准误差 68

5.3.1 自举法的标准误差 68

5.3.2 大折刀法的标准误差 70

5.4 多重借补的介绍 73

5.5 再抽样方法和多重借补的比较 77

问题 78

第二部分 用于缺失数据分析的基于似然的方法

第六章 基于似然函数的推断理论	83
6.1 完全数据基于似然的估计的回顾	83
6.1.1 极大似然估计	83
6.1.2 贝叶斯估计基础	89
6.1.3 大样本的极大似然和贝叶斯推断	90
6.1.4 基于完整后验分布的贝叶斯推断	96
6.1.5 后验分布的模拟抽取值	99
6.2 不完全数据基于似然的推断	101
6.3 极大似然以外通常有缺陷的方法:对参数和缺失数据极大化	106
6.3.1 方法	106
6.3.2 背景	106
6.3.3 例子	107
6.4 对粗化数据的似然理论问题	109 111
第七章 因子化似然方法,忽略缺失数据机制	114
7.1 引言	114
7.2 具有一个变量不响应的二元正态数据:ML 估计(极大似然估计)	115
7.2.1 ML 估计	115
7.2.2 大样本协差阵	119
7.3 二元正态单调数据:小样本推断	120
7.4 两个以上变量的单调数据	123
7.4.1 有一个正态变量不响应的多变量数据	123
7.4.2 一般单调模式的似然的因子化	124
7.4.3 单调正态数据用扫描算子的 ML 计算	125
7.4.4 单调正态数据用扫描算子的贝叶斯计算	133
7.5 对特殊的非单调模式的因子化问题	134 138
第八章 缺失数据一般模式的极大似然,可忽略不响应的介绍和理论	141
8.1 另一种可选用的计算策略	141

8.2 EM算法的介绍	143
8.3 EM的E步和M步	144
8.4 EM算法的理论	148
8.4.1 收敛性质	148
8.4.2 指数族的EM	150
8.4.3 EM的收敛的速率	152
8.5 EM的推广	154
8.5.1 ECM算法	154
8.5.2 ECME和AECM算法	158
8.5.3 PX—EM算法	159
8.6 混合极大化方法	161
问题	162

第九章 基于极大似然估计的大样本推断 164

9.1 基于信息阵的标准误差	164
9.2 无需计算已观测信息矩阵的估计并求逆,产生标准误差	165
9.2.1 补充的EM算法	165
9.2.2 自举观测到的数据	169
9.2.3 其他的大样本方法	170
9.2.4 贝叶斯方法的后验标准误差	171
问题	171

第十章 贝叶斯和多重借补 173

10.1 贝叶斯迭代模拟方法	173
10.1.1 数据扩张(DA)	173
10.1.2 Gibbs抽样	176
10.1.3 评价迭代模拟的收敛性	178
10.1.4 一些其他的模拟方法	180
10.2 多重借补	181
10.2.1 基于小数量抽取的大样本贝叶斯后验均值和方差的近似	181
10.2.2 使用检验统计量的近似	183
10.2.3 创建多重借补的其他方法	185
10.2.4 在借补和分析中不同模型的使用	188
问题	189

第三部分 不完全数据分析的基于似然的方法:一些例子

第十一章	多元正态的例子,可忽略缺失数据机制	193
11.1	引言	193
11.2	正态下有缺失数据时均值向量和协差阵的推断	193
11.2.1	不完全多元正态样本的 EM 算法	193
11.2.2	估计 $(\theta - \hat{\theta})$ 的渐近协差阵	195
11.2.3	借助于数据扩张的正态模型的贝叶斯推断	196
11.3	有限制协差阵的估计	201
11.4	多元线性回归	206
11.4.1	限于相依变量有缺失值的线性回归	206
11.4.2	有缺失数据的更一般的回归问题	207
11.5	一个一般的有缺失数据的重复度量模型	210
11.6	时间序列模型	213
11.6.1	引言	213
11.6.2	有缺失值的一元时间序列的自回归模型	213
11.6.3	卡尔曼滤波模型	216
	问题	218
第十二章	稳健估计	220
12.1	引言	220
12.2	一元样本的稳健估计	220
12.3	均值和协差阵的稳健估计	222
12.3.1	多元完全数据	222
12.3.2	数据有缺失值时均值和协差阵的稳健估计	223
12.3.3	自适应稳健多元估计	224
12.3.4	t 模型的贝叶斯推断	226
12.4	t 模型的进一步扩展	227
	问题	230
第十三章	未完全分类的列联表模型,忽略缺失数据机制	231
13.1	引言	231
13.2	单调多项数据的因子化似然	232
13.2.1	引言	232
13.2.2	单调模式的 EM 估计	233
13.2.3	估计的精度	239

13.3 有一般缺失数据模式的多项分布样本的 ML 和贝叶斯估计	241
13.4 不完全分类列联表的对数线性模型	244
13.4.1 完全数据的情形	244
13.4.2 未完全分类表的对数线性模型	248
13.4.3 未完全分类数据的拟合优度检验	251
问题	252

第十四章 有缺失值的正态和非正态混合数据,可忽略缺失数据机制 254

14.1 引言	254
14.2 一般的位置模型	254
14.2.1 完全数据模型和参数估计	254
14.2.2 有缺失值的 ML 估计	256
14.2.3 E 步计算的详情	257
14.2.4 无限制一般的位置模型的贝叶斯计算	259
14.3 有参数制约的一般的位置模型	261
14.3.1 引言	261
14.3.2 对小格均值的有限制的模型	261
14.3.3 小格概率的对数线性模型	261
14.3.4 参数有限制时 14.2.2 和 14.2.3 节的算法的修正	264
14.3.5 当分类变量比连续变量有更多观测时的简化	266
14.4 连续和分类变量混合的回归问题	267
14.4.1 有缺失连续变量或分类变量的正态线性回归	267
14.4.2 有缺失连续或分类变量的 Logistic 回归	268
14.5 一般的位置模型的进一步扩展	269
问题	269

第十五章 不可忽略缺失数据模型 271

15.1 引言	271
15.2 不可忽略模型的似然理论	273
15.3 具有已知不可忽略缺失数据机制的模型:分组的和归并的数据	274
15.4 正态的选择模型	279
15.5 正态模式混合模型	284

15.5.1 一元正态模式混合模型	284
15.5.2 借助参数限制识别的二元正态模式混合模型	288
15.6 正态重复测量数据的不可忽略模型	292
15.7 分类数据的不可忽略模型	295
问题	300

参考文献	303
------	-----



第一部分

回顾与基本方法

