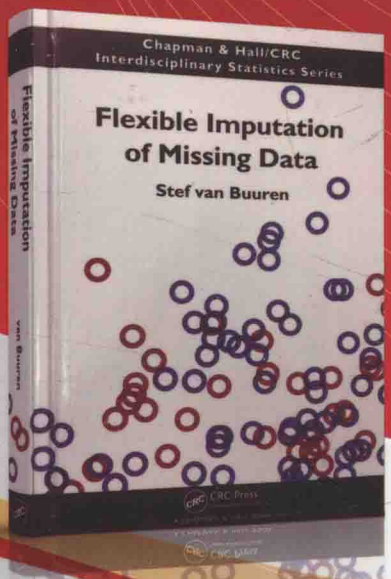
国外实用统计丛书

缺失数据的灵活填补方法

（英文影印注释版）

Flexible Imputation of Missing Data



[荷] 史蒂夫·范·布伦（Stef Van Buuren） 著

CRC Press
Taylor & Francis Group

机械工业出版社
CHINA MACHINE PRESS

国外实用统计丛书

缺失数据的灵活填补方法

（英文影印注释版）

[荷] 史蒂夫·范·布伦（Stef Van Buuren）著
刘 俊 夏爱生 索文莉 鞠 涛 注释

机械工业出版社

本书分为三部分：第Ⅰ部分基础篇、第Ⅱ部分案例分析、第Ⅲ部分延伸，共10章。作者结合众多实验中的例子，探讨如何解决缺失数据这一广泛存在于各个领域之中的问题，深入地讨论了解决这类问题的方法，并分析了每种方法的适用范围和优缺点。书中的算法结合统计软件来实现，主要内容包括多元缺失填补、单变量缺失数据、多变量缺失数据、数据填补实践、填补数据分析、测量、选择、纵向数据、结论等。

本书可作为高等院校统计学专业的本科高年级学生以及研究生用书，也可作为与统计学专业相关的科研人员的参考书。

图书在版编目（CIP）数据

缺失数据的灵活填补方法：英文影印注释版：英文/（荷）史蒂夫·范·布伦著；（中国）刘俊等注释. —北京：机械工业出版社，2017.10
（国外实用统计丛书）
书名原文：Flexible Imputation of Missing Data
ISBN 978-7-111-58416-2

Ⅰ. ①缺… Ⅱ. ①史… ②刘… Ⅲ. ①统计数据－数据处理－高等学校－教材－英文 Ⅳ. ①O212

中国版本图书馆CIP数据核字（2017）第270099号

机械工业出版社（北京市百万庄大街22号 邮政编码100037）
策划编辑：汤 嘉 责任编辑：汤 嘉 韩效杰
责任印制：张 博
三河市宏达印刷有限公司印刷
2018年2月第1版第1次印刷
184mm×242mm · 22印张 · 9插页 · 489千字
标准书号：ISBN 978-7-111-58416-2
定价：78.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务	网络服务
服务咨询热线：010-88361066	机工官网：www.cmpbook.com
读者购书热线：010-68326294	机工官博：weibo.com/cmp1952
读者购书热线：010-88379203	金 书 网：www.golden-book.com
封面无防伪标均为盗版	教育服务网：www.cmpedu.com

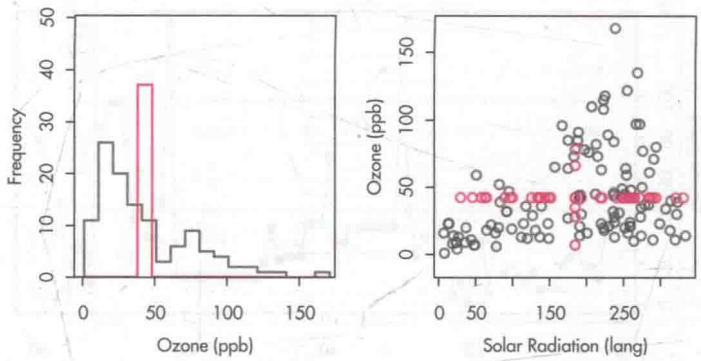


Figure 1.1: Mean imputation of `Ozone`. Gray indicates the observed data, red indicates the imputed values.

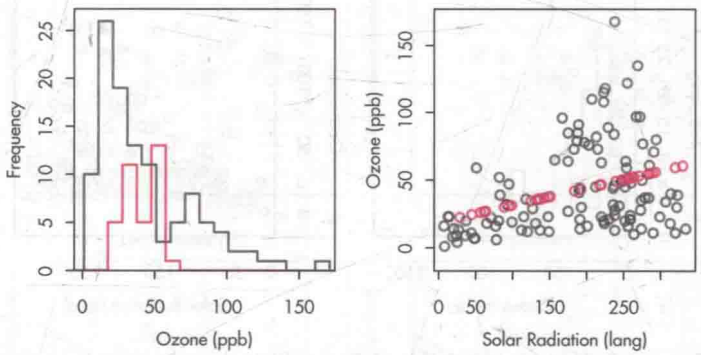


Figure 1.2: Regression imputation: Imputing `Ozone` from the regression line.

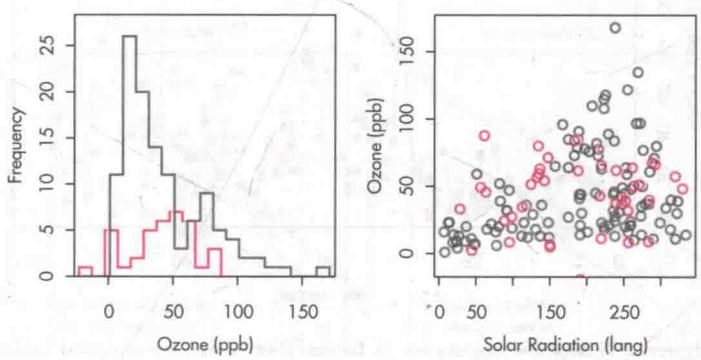


Figure 1.3: Stochastic regression imputation of `Ozone`.

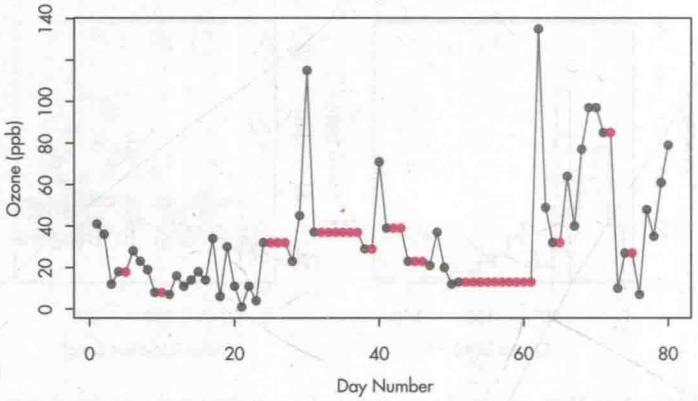


Figure 1.4: Imputation of `Ozone` by last observation carried forward (LOCF).

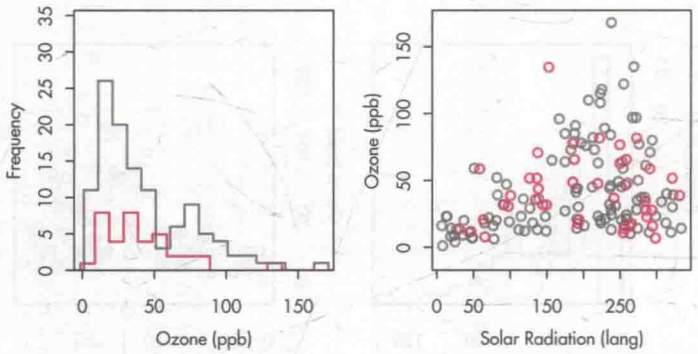


Figure 1.6: Multiple imputation of `Ozone`. Plotted are the imputed values from the first imputation.

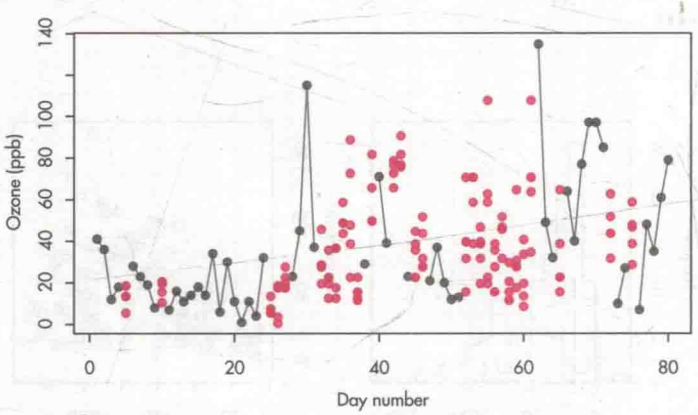


Figure 1.7: Multiple imputation of `Ozone`. Plotted are the observed values (in gray) and the multiply imputed values (in red). One red dot at (61,168) is not plotted.

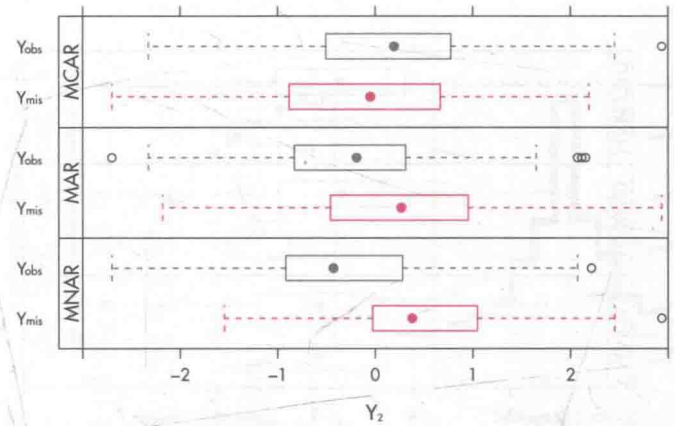


Figure 2.2: Distribution of Y_{obs} and Y_{mis} under three missing data models.

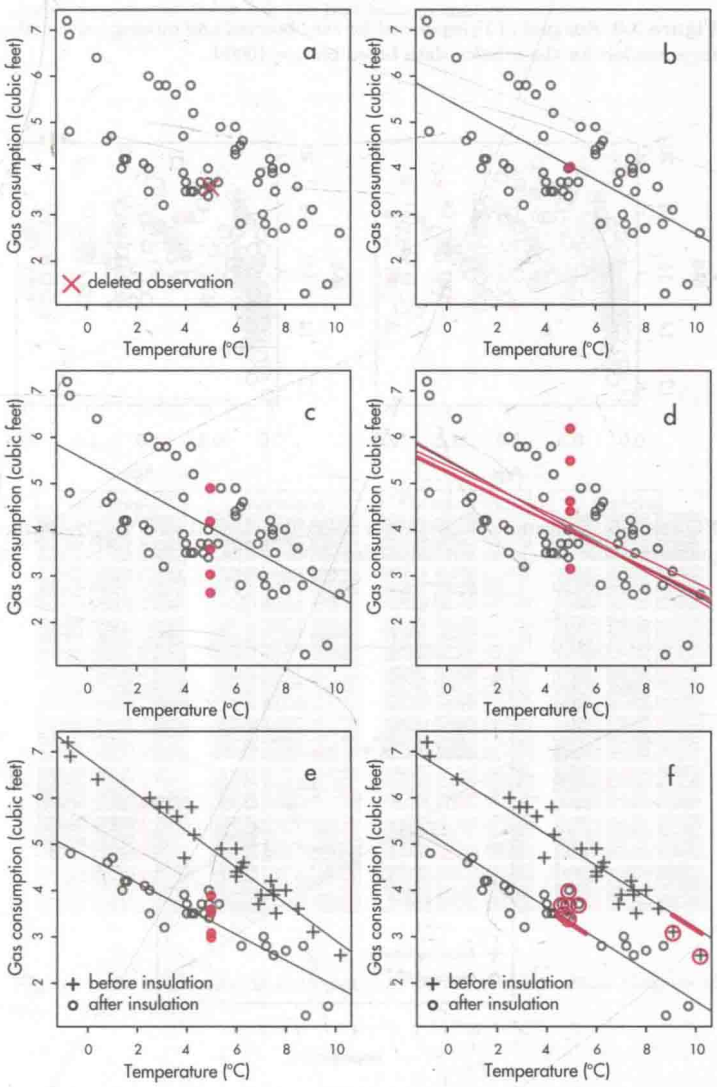


Figure 3.1: Five ways to impute missing gas consumption for a temperature of 5°C: (a) no imputation; (b) predict; (c) predict + noise; (d) predict + noise + parameter uncertainty; (e) two predictors; (f) drawing from observed data.

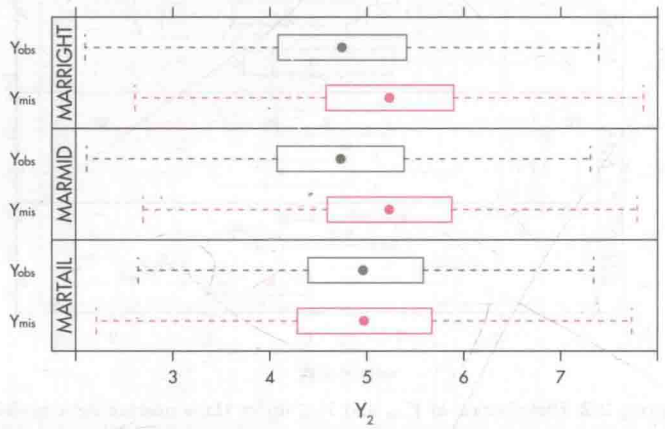


Figure 3.3: Box plot of Y_2 separated for the observed and missing parts under three models for the missing data based on $n = 10000$.

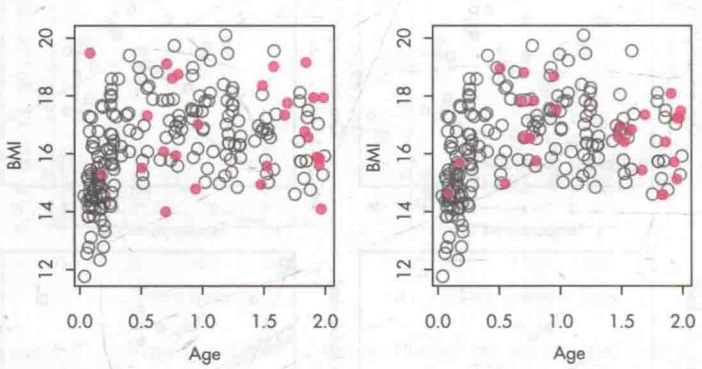


Figure 3.6: Robustness of predictive mean matching (right) relative to imputation under the linear normal model (left).

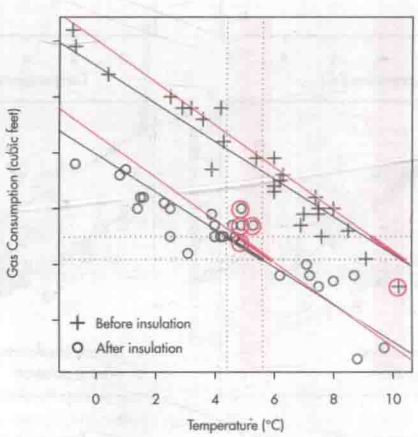


Figure 3.7: Selection of candidate donors in predictive mean matching with the stochastic matching distance.

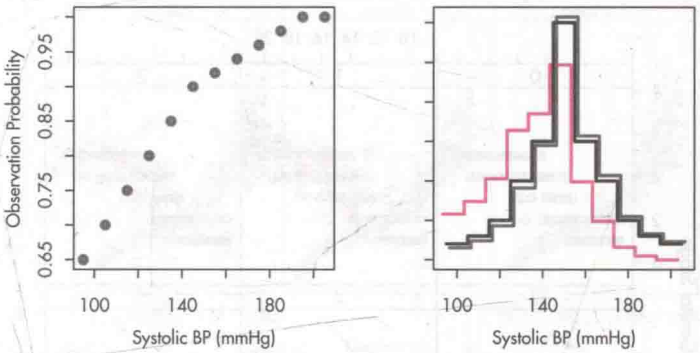


Figure 3.9: Graphic representation of the response mechanism for systolic bloood pressure in Table 3.5. See text for explanation.

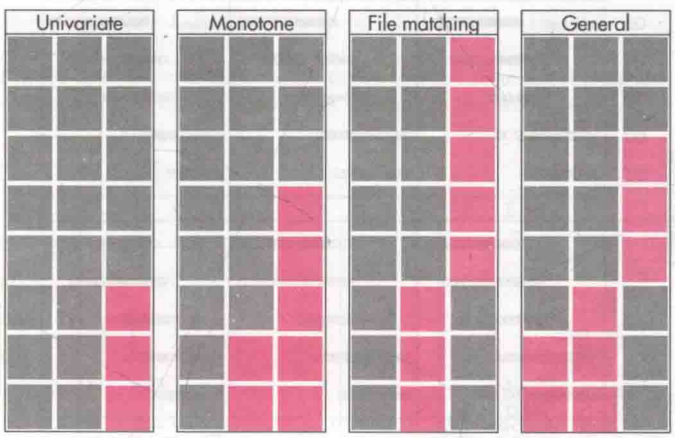


Figure 4.1: Some missing data patterns in multivariate data. Gray is observed, red is missing.

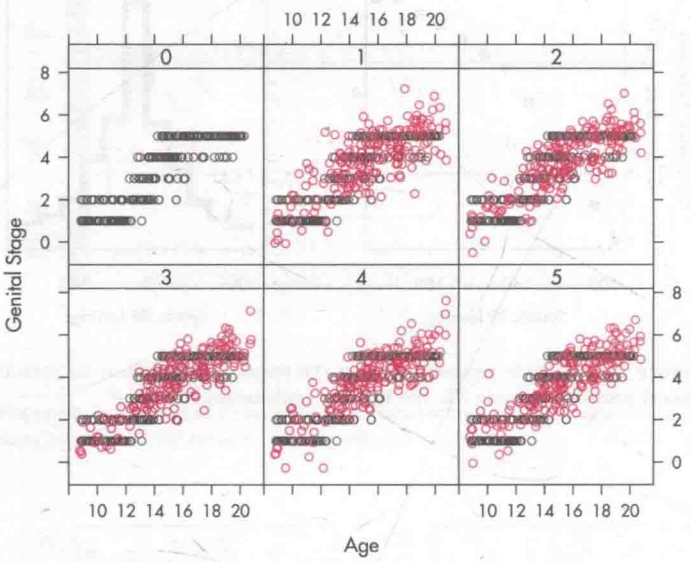


Figure 4.4: Joint modeling: Imputed data for genital development (Tanner stages G1–G5) under the multivariate normal model. The panels are labeled by the imputation numbers 0–5, where 0 is the observed data and 1–5 are five multiply imputed datasets.

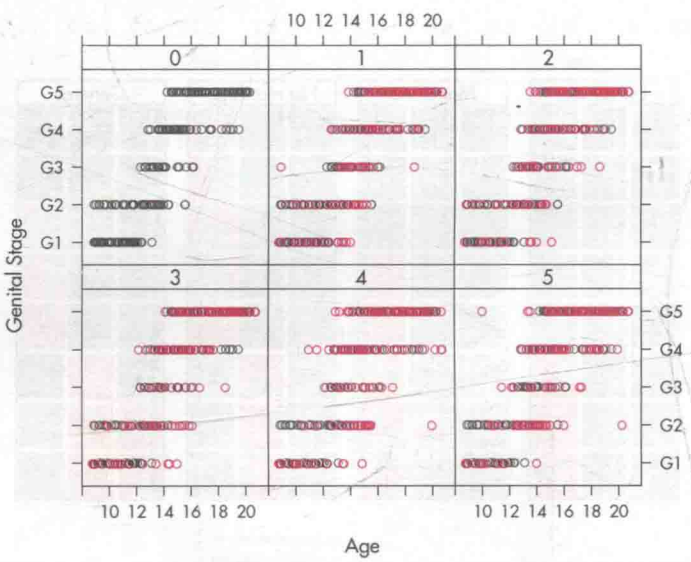


Figure 4.5: Fully conditional specification: Imputed data of genital development (Tanner stages G1–G5) under the proportional odds model.

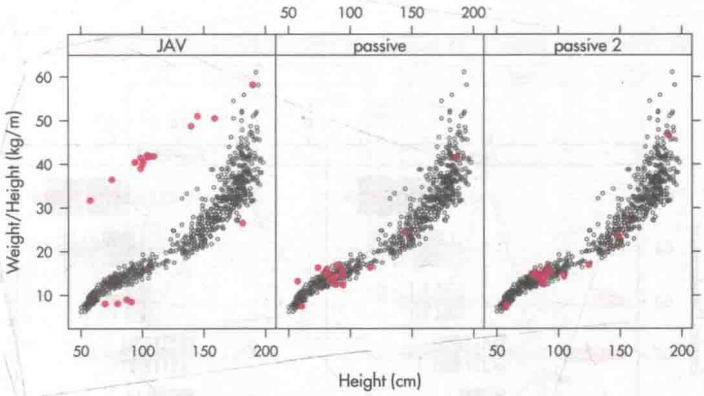


Figure 5.1: Three different imputation models to impute weight/height ratio (`whr`). The relation between `whr` and height (`hgt`) is not respected under "just another variable" (JAV). Both passive methods yield imputations that are close to the observed data. "Passive 2" does not allow for models in which `whr` and `bmi` are simultaneous predictors.

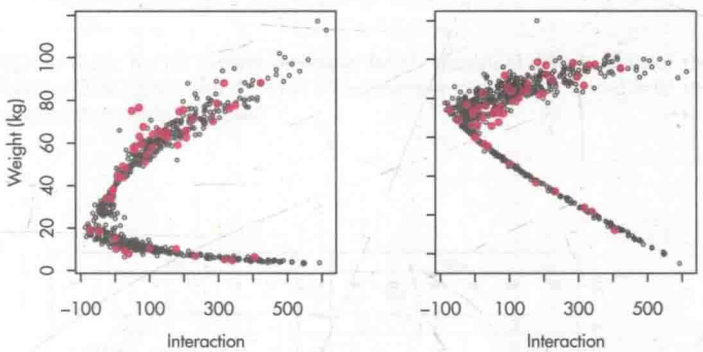


Figure 5.2: The relation between the interaction term `wgt.hc` (on the horizontal axes) and its components `wgt` and `hc` (on the vertical axes).

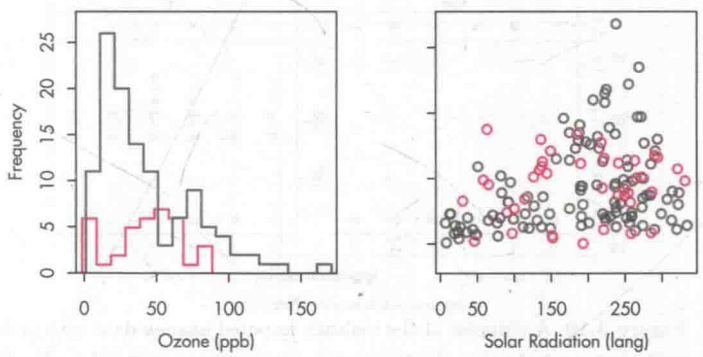


Figure 5.3: Stochastic regression imputation of `Ozone`, where the imputed values are restricted to the range 1–200. Compare to Figure 1.3.

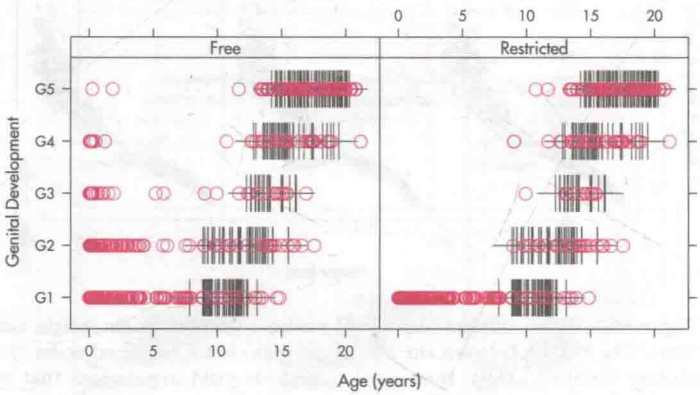


Figure 5.4: Genital development of Dutch boys by age. The "free" solution does not constrain the imputations, whereas the "restricted" solution requires all imputations below the age of 8 years to be at the lowest category.

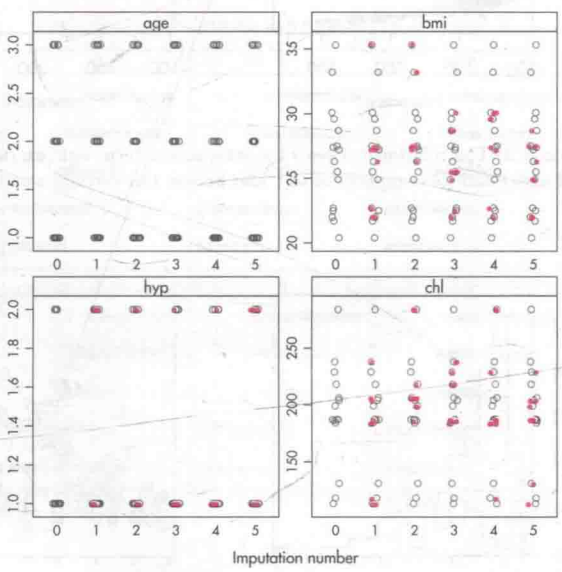


Figure 5.10: A stripplot of the multiply imputed nhanes data with $m = 5$.

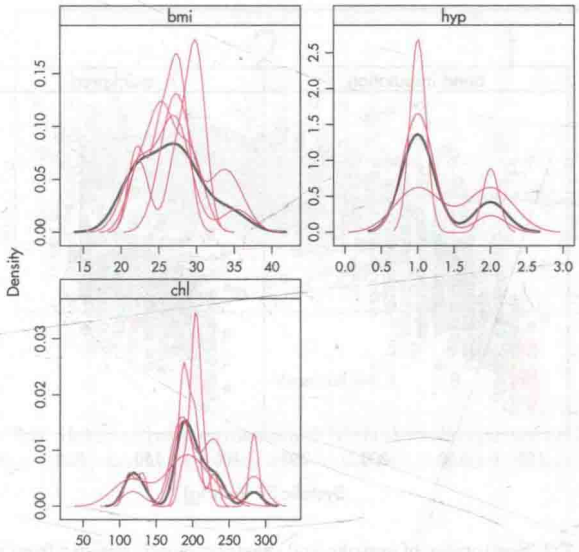


Figure 5.11: Kernel density estimates for the marginal distributions of the observed data (gray) and the $m = 5$ densities per variable calculated from the imputed data (thin red lines).

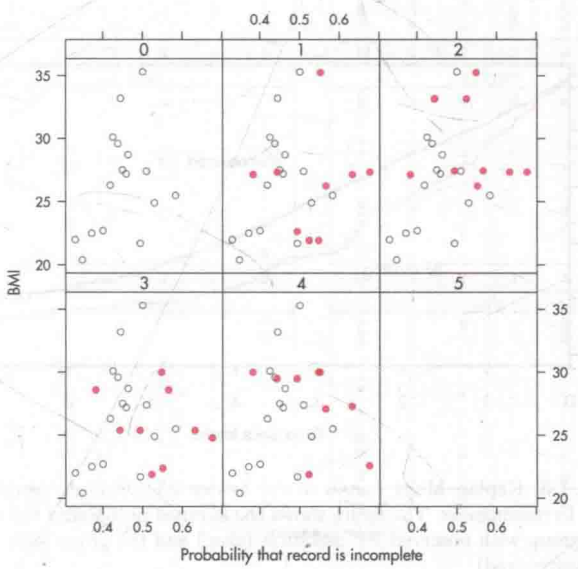


Figure 5.12: BMI against missingness probability for observed and imputed values.

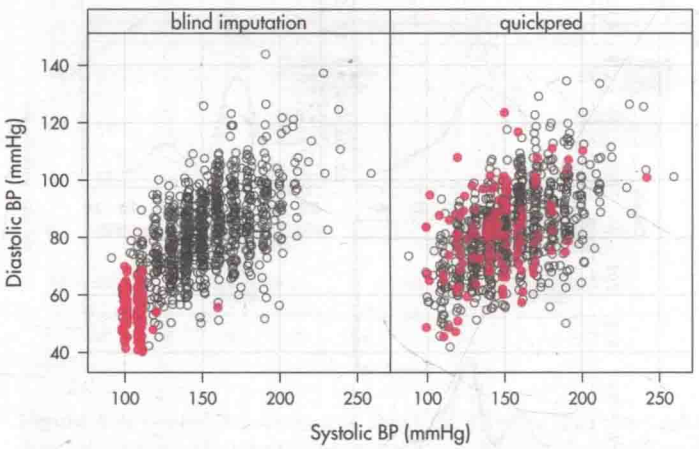


Figure 7.2: Scatterplot of systolic and diastolic blood pressure from the first imputation. The left-hand-side plot was obtained after just running `mice()` on the data without any data screening. The right-hand-side plot is the result after cleaning the data and setting up the predictor matrix with `quickpred()`. Leiden 85+ Cohort data.

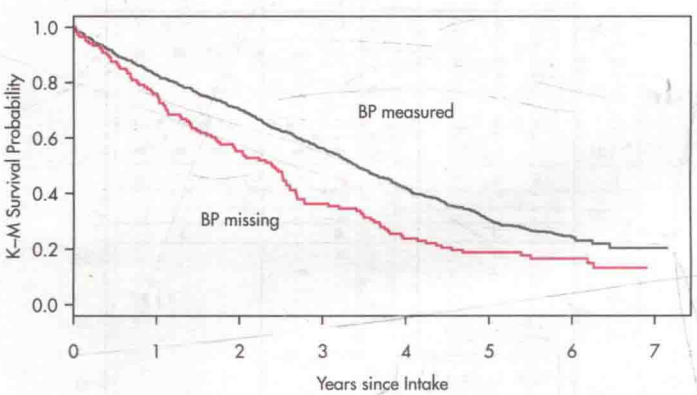


Figure 7.3: Kaplan–Meier curves of the Leiden 85+ Cohort, stratified according to missingness. The figure shows the survival probability since intake for the group with observed BP measures (gray) and the group with missing BP measures (red).

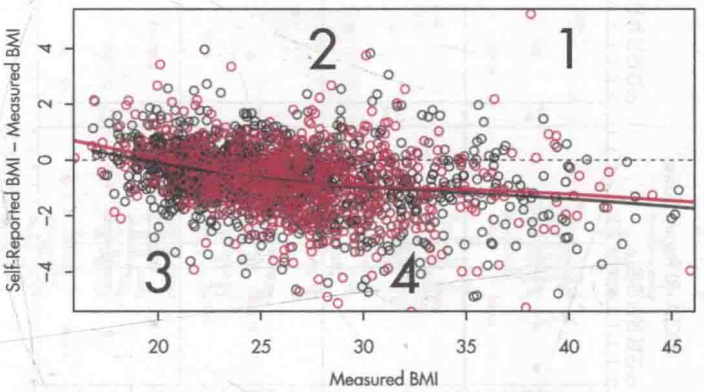


Figure 7.6: Relation between measured BMI and self-reported BMI in the calibration (gray) and survey (red) data in the first imputed dataset.

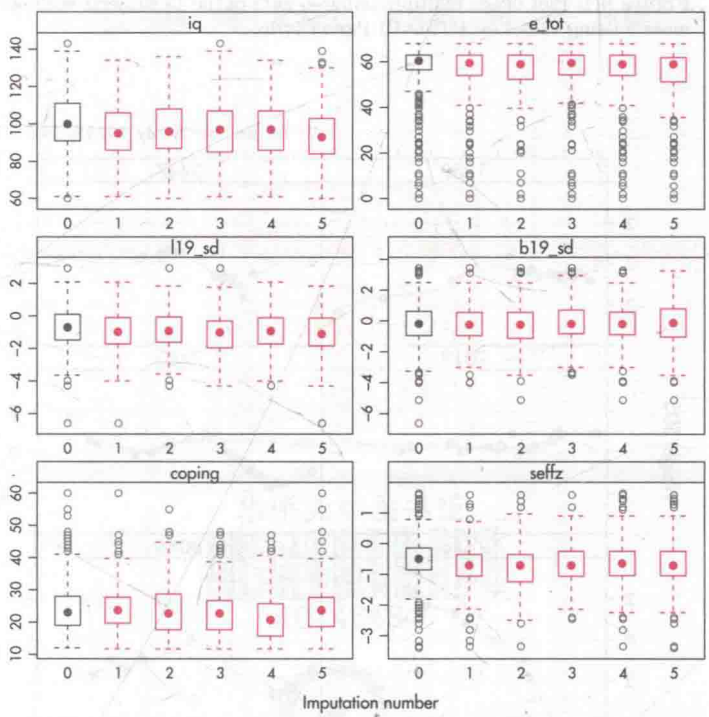


Figure 8.3: Distributions (observed and imputed values) of six outcome variables at 19 years in the POPS study.

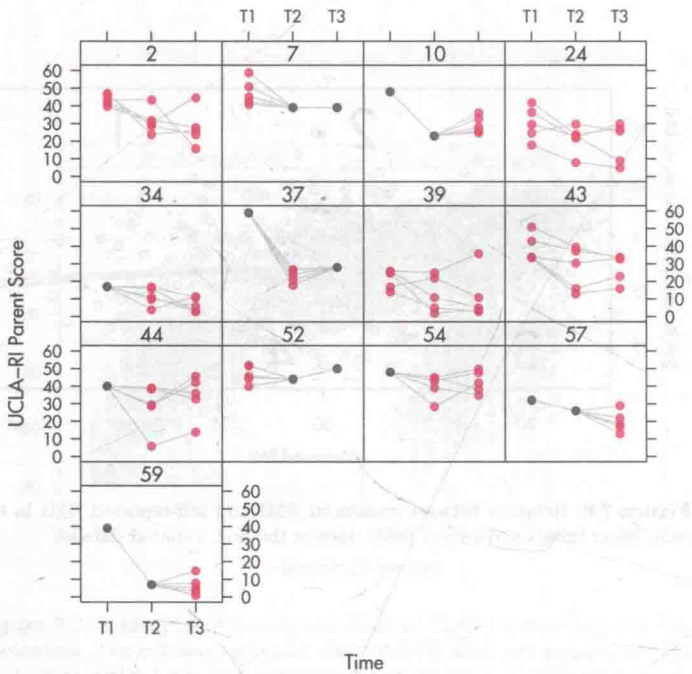


Figure 9.1: Plot of the multiply imputed data of the 13 subjects with one or more missing values on PTSD-RI Parent form.

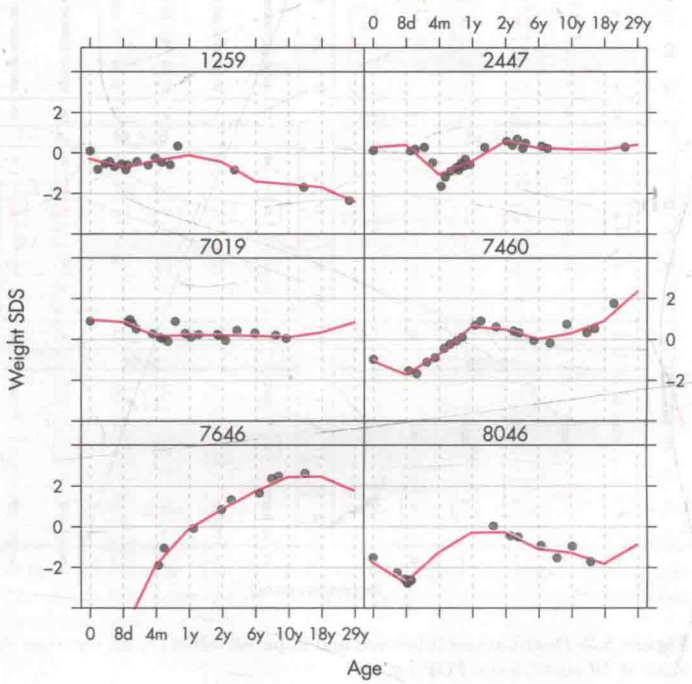


Figure 9.5: Broken stick trajectories for Weight SDS from six selected individuals from the Terneuzen cohort.

序（译）

史蒂夫·范·布伦所著的关于多重估算法的著作我很喜欢和欣赏。第一个原因是在不被公众广泛接受的情况下，又一本研究关于多重估算的书籍出版标志着这个方向的研究日渐成熟，至少我是这样认为的。史蒂夫在2.1.2节的陈述是对的：创造多个版本的主意在当时（20世纪70年代末）肯定令人无法接受。

用从分布获得的虚拟值代替估算的最佳值摒弃了以前固有的方法。曾几何时，多重估算遭到许多较传统的统计学家们的质疑，被认为是“愚蠢的”，有人认为由于存储需求，它的效率低下，也有人认为由于计算量的要求，它的成本过高。

有一些人已预知了计算存储（a）、计算速度（b）和灵活性的变化。（我刚刚用不到60美元买到一个小手指大小的64GB闪存盘，然而在几十年前，我得花2500多美元才能买到一个重10千克、比鞋盒还大且内存仅为120KB的硬盘）。

基于过去计算方法的限制，开发将来才能实现的新的计算方法显然是不合适的。多重估算得以发展完全得益于新一代有远见的统计学家们，这其中包括许多见多识广的同事、校友。

喜欢这本书的第二个原因是，我个人对作者的喜爱和关注。正如他所说，我们通过杨·范·瑞吉克沃塞勒（Jan Van Rijckevorsel）介绍第一次见面时，他当时是那里一名年轻有为、充满热情的研究人员，关于缺失数据的处理，他所知甚少。但是考虑到他从早期研究在线仿真器到现在多年来的进步，他已经成长为一名独立的研究员，并为多重估算的发展做出了重要的贡献。

这本书呈现出了一种实用的、直截了当的多重估算的应用方法。我尤其喜欢史蒂夫对图形显示的运用，这对于实践中补充多重估算法的一般有效性的理论探讨十分必要。正如我曾经说过的以及书中所提及的，“对于缺失数据处理，并非多重估算法有多好，而是其他方法太糟糕了”。

唐纳德 B. 鲁宾（Donald B.Rubin）

前言（译）

我们时常被缺失数据的情况所困扰。统计分析中由缺失数据所带来的问题长期被掩盖，现在这种情况正在慢慢结束。近十年间，处理缺失数据的技术迅速得到补充和发展。本书主要介绍一种方法：多重插补。

多重插补是统计科学领域重要的思想之一。这种技术简便、巧妙而且强大。说它简便是因为它填补了由似是而非的数据造成的漏洞，说它巧妙是因为未知数据的不确定性被数据本身所标记，说它强大是因为它可以解决那些被掩饰的数据缺失问题。

在近二十年的时间里，我已经将多重插补应用到了更广泛的研究领域中。我相信多重插补进入统计学主流的时机已经成熟。当今计算机和软件技术已能够充分满足计算的需要。我们所欠缺的是关于介绍这些基本思想及这些思想该如何应用的书。我希望这本书能够弥补这个欠缺。

本书正文的阅读要求读者通晓统计学基础概念和多元统计方法。本书特别为如下两类读者而设计：

* 社会和健康科学领域的（生物）统计学家、流行病学家等；

* 不称呼自己为统计学家，但拥有必要的技能来理解规则并使用一些统计方法的独立的研究者。

在编写本书的时候，我尽量避免数学和技术上的细节，出现公式的地方往往辅之以图表，并用可视化的陈述来解释该公式。我希望读者朋友们可以较少去关注理论基础，而更多去抓住宏观的总体上的思路。偏技术的内容在本书中标记了黑色桃心，这在第一次阅读时可以暂时跳过。

我在乌得勒支大学采用了本书中的一些章节来教授研究生插补技术的课程，主要的基础内容体现在 1~4 章，大约要花费十个小时来讲授这些材料，中间留出时间可以让学生们完成书中的练习题。

本书采用了大量唐纳德·鲁宾（Donald Rubin）的理念，他是多元插补这个学科的奠基人之一。我非常有幸在很多场合与他见面、讨论和工作。他富于逻辑的设想和貌似简单的想法是我的灵感的极佳的来源。同时要感谢