研究生教学用书

纵向数据分析

Analysis of Longitudinal Data

王友乾　付利亚　徐建文　编著

高等教育出版社

研究生教学用书

纵向数据分析

ZONGXIANG SHUJU FENXI

Analysis of Longitudinal Data

王友乾　付利亚　徐建文　编著

高等教育出版社·北京

内容简介

本书以实例为背景，系统阐述了纵向数据分析中边际模型估计参数的方法、参数估计的渐近性质、模型中相关矩阵和协变量的选择及其在实际数据中的应用。全书共分八章，内容包括纵向数据的背景，线性模型，广义线性模型，边际模型，参数估计的协方差矩阵估计，模型选择，纵向数据的秩的统计推断，拓展话题。

本书可作为高等学校统计学类、生物医药类、环境科学类等相关专业高年级本科生或研究生的教材，也可作为相关领域的科技工作者的参考用书。

图书在版编目（CIP）数据

纵向数据分析 / 王友乾，付利亚，徐建文编著．--
北京：高等教育出版社，2015.12
ISBN 978-7-04-043889-5

Ⅰ.①纵… Ⅱ.①王… ②付… ③徐… Ⅲ.①统计数据－统计分析－高等学校－教材 Ⅳ.①O212.1

中国版本图书馆 CIP 数据核字（2015）第 223973 号

策划编辑	张晓丽	责任编辑	张晓丽	特约编辑	胡 宇	封面设计	张卫青
版式设计	王艳红	插图绘制	杜晓丹	责任校对	张小镝	责任印制	刘思涵

出版发行	高等教育出版社	咨询电话	400-810-0598
社　　址	北京市西城区德外大街 4 号	网　　址	http://www.hep.edu.cn
邮政编码	100120		http://www.hep.com.cn
印　　刷	北京凌奇印刷有限责任公司	网上订购	http://www.landraco.com
开　　本	787 mm×960 mm　1/16		http://www.landraco.com.cn
印　　张	10.75	版　　次	2015 年 12 月第 1 版
字　　数	190 千字	印　　次	2015 年 12 月第 1 次印刷
购书热线	010-58581118	定　　价	19.00 元

物 料 号　43889-00

序

本书是由我的好朋友，著名统计学家王友乾教授带领他的两个学生付利亚博士、徐建文博士撰写的。王友乾教授曾先后就读于浙江大学、北京大学和牛津大学，曾在哈佛大学、新加坡国立大学、澳大利亚联邦科工组织和昆士兰大学从事统计方面的教学与研究。他在纵向数据分析领域有着很高的造诣。

纵向数据分析起源于临床试验中的试验设计与数据分析。为检验一种新药的医疗效果，医药学家需要对参与临床试验的试验者进行长期的跟踪观察，然后利用严格的数理统计分析方法来确定新药的疗效。它的特点是同时观察若干个试验者，且在每个试验者身上测量若干个数据。纵向数据的结构不同于一般回归分析数据，我们不能假定来自同一个试验者的不同观测数据是统计独立的。这种特殊性决定了纵向数据分析的特点和难点，也给纵向数据的建模与分析带来了一定的困难。由于这个特点，统计中常用的似然方法很难直接用来进行数据分析，因此需要将新的方法引入到纵向数据分析中来，例如约束极大似然估计、拟似然估计、广义估计方程等。

自 20 世纪 40 年代以来，纵向数据分析已经逐步成为一门十分重要的统计分支。除了临床医学以外，纵向数据分析在许多领域都有应用，例如生物生态平衡研究，森林植被研究，经济金融管理、军事侦测等。近年来，纵向数据分析的研究和应用得到了飞速的发展，但是国内还没有一本系统讲解纵向数据分析理论及方法的中文教材。本书理论和实际数据相结合，通俗易懂，且面向前沿，重点介绍了纵向数据边际建模及其相关问题。该书的出版将为研究生的统计学习和实际工作者在纵向数据分析方面的学习与应用提供必要的指引和有利的帮助。

白志东

2014 年 12 月于长春

前　言

近几年来, 纵向数据的统计分析引起广大统计学家和实际应用者的关注, 主要因为纵向数据在生物、医学、社会经济学、心理、教育、环境科学等领域存在着广泛应用。

在国内外, 很多高等院校已经将纵向数据分析列入数学系、统计系、生物系等高年级本科生、硕士生或者博士生的学位课或选修课。一系列有关纵向数据分析的英文专著也相继出版, 但国内鲜有介绍纵向数据的教材, 本书是为了方便国内高年级学生和相关研究者的阅读和参考而编写的教材或参考书。

全书共分八章, 第一章主要是通过实际例子介绍了纵向数据的定义及其特点, 使读者对纵向数据的实际背景有一定的了解, 有助于理解后面引进的模型和方法。第二章由大家熟知的最简单的线性模型过渡到纵向数据下的线性模型, 主要介绍纵向数据下线性模型参数的加权最小二乘估计、极大似然估计和带约束的极大似然估计, 最后一节通过实例分析使读者对讲述的方法有更深入透彻的理解。第三章将线性模型推广到了广义线性模型, 并进一步推广到拟似然方法, 为第四章引入广义估计方程方法作铺垫。第四章主要讲述了估计边际模型中的回归系数、相关系数和方差参数比较经典的估计方法以及估计的大样本性质。后四章内容是对第四章的进一步延伸和推广。其中, 第五章主要介绍了回归系数估计的协方差矩阵的一些估计方法, 第六章主要介绍了边际模型中选择协变量和相关矩阵的各种经典准则, 第七章介绍了纵向数据中基于秩的统计推断。最后一章介绍了广义估计方法的一些最新推广, 以及可以对不连续估计函数进行光滑的诱导平滑方法。

本书的第一作者曾在澳大利亚昆士兰大学讲授过本书的部分内容。高年级本科生和硕士生可以只讲述前四章内容, 后四章的每一章均可以作为一个独立专题来研究, 感兴趣的读者可以自学相关内容。为了方便读者对本书讲述方法的理解和运用, 作者在实例分析中给出了统计软件 R 下的运行结果及其结果的解读, 相关程序和数据可在 http://gr.xjtu.edu.cn/web/fuliya/8 下载。

本书的写作得到国家自然科学基金 (No.11201365、No.11301408 和 No.11201505)、教育部新教师博士点基金 (2012020112005) 以及中央高校基金 (CQDXWL-2013-Z009)、重庆市基础与前沿研究计划项目一般项目 (cstc2013jcyjA00001) 的资助,

另外, 本书的出版得到高等教育出版社张晓丽女士的支持和关心, 编者愿借此机会向他们表示诚挚的谢意。

本书由王友乾教授担任主编, 负责统稿。其中, 第一章、第二章的 2.3 至 2.5 节、第四章的 4.2 至 4.4 节、第五章的 5.2 节、第六章的 6.1.3 节和 6.2 节、第七章以及第八章的 8.1.2 节、8.2 至 8.3 节由付利亚执笔, 第二章的 2.1 和 2.2 节、第三章、第四章的 4.1 节、第五章的 5.1 节、第六章的 6.1.1 节和 6.1.2 节以及第八章的 8.1.1 节由徐建文执笔。诚恳希望国内同行及广大读者指出本书的不足和提供宝贵的意见, 借此机会可以相互学习相互改进。

编者

2014 年 12 月 22 日

王友乾[①]　昆士兰科技大学　you-gan.wang@qut.edu.au
付利亚　西安交通大学　fuliya@mail.xjtu.edu.cn
徐建文　重庆大学　xjw@cqu.edu.cn

① 王友乾, 现任澳大利亚昆士兰科技大学终身教授、博士生导师。1986 年毕业于浙江大学数学系, 随后到北京大学统计系开始硕士研究生学习, 1988 年进入牛津大学学习, 1991 年获牛津大学统计学博士学位。曾任澳大利亚昆士兰大学统计系首席教授、自然资源和应用数学研究中心主任, 澳大利亚联邦科工组 (CSIRO) 资深研究员和首席科学家, 哈佛大学生物统计学副教授和新加坡国立大学生物统计学副教授。主要研究领域有纵向数据分析、模型选择和优化、稳健估计及推断、水资源和水文学统计模型、渔业资源评估与管理等。

先后在统计学顶尖杂志 *Annals of Statistics*、*Journal of American Statistical Association*、*Biometrika*、*Biometrics* 等期刊发表 SCI 论文 120 多篇。是国际统计协会会员, IMS 永久会员, 并先后担任 *Electronic Journal of Statistics* 和 *Biometrics* 副主编, *Environmental Modelling and Assessment* 编委。

目　录

第一章 纵向数据的背景

在本章, 我们主要介绍什么是纵向数据, 以及几个实际例子.

1.1 什么是纵向数据

纵向数据 (longitudinal data) 是指对一系列试验个体随着时间的演变进行跟踪测量得到的数据. 更确切地说, 假设在一项研究中有 m 个个体, 对每个个体随着时间的推移进行测量, 对第 i 个试验个体, 在时刻 $t_{i1} < t_{i2} < \cdots < t_{in_i}$ 测量得到的数据为 $Y_{i1}, Y_{i2}, \cdots, Y_{in_i}$, $i = 1, \cdots, m$, 则 $\{Y_{ik} : 1 \leqslant k \leqslant n_i, 1 \leqslant i \leqslant m\}$ 就是纵向数据. 在纵向数据中, 试验个体通常称为被试 (subject), 可以是医学研究中的患者, 林业研究中的树木, 生物学研究中的动物, 也可以是水和空气质量研究中收集数据的位点.

下面我们给出一个简单的纵向数据集列表. 不难看出, 如果固定试验个体 i, 则每一行就是一个时间序列, 但一般来说纵向数据的每个个体的时间序列都很短 ($n_i = 2, 3, 4$ 等); 如果固定时间, 则每一列就是一组截面数据. 因此, 纵向数据是由一批短时序的时间序列构成, 是将截面数据和时间序列数据结合在一起, 其兼有时间序列与多元分析的特性. 但是, 纵向数据不同于传统意义上的时间序列, 因为通常情况下研究的时间序列是针对一个个体 (例如某一个地区的每年的降雨量), 而且 n_i 比较大, 它也不同于传统的截面数据, 因为截面数据是对每个个体只测量一次, 而纵向数据是对每个个体测量多次. 因此, 相对于截面数据和时间序列, 纵向数据分析更为复杂. 在纵向数据中, 如果 $n_1 = n_2 = \cdots = n_m$, 我们称其为平衡数据, 否则称为非平衡数据.

在分析纵向数据时, 有一个基本假设: 不同个体的测量值之间是相互独立的,

i	t_{i1}	t_{i2}	t_{i3}	$\cdots$	t_{i,n_i-2}	t_{i,n_i-1}	t_{in_i}
1	Y_{11}	Y_{12}	Y_{13}	$\cdots$	Y_{1,n_1-2}	Y_{1,n_1-1}	Y_{1n_1}
2	Y_{21}	Y_{22}	Y_{23}	$\cdots$	Y_{2,n_2-2}	Y_{2,n_2-1}	Y_{2n_2}
3	Y_{31}	Y_{32}	Y_{33}	$\cdots$	Y_{3,n_3-2}	Y_{3,n_3-1}	Y_{3n_3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
m	Y_{m1}	Y_{m2}	Y_{m3}	$\cdots$	Y_{m,n_m-2}	Y_{m,n_m-1}	Y_{mn_m}

但来自于同一个个体的测量值之间是相关的, 即 $(Y_{11}, Y_{12}, \cdots, Y_{1n_1}), \cdots, (Y_{m1}, Y_{m2}, \cdots, Y_{mn_m})$ 之间相互独立, 但对于每个 i, $Y_{i1}, Y_{i2}, \cdots, Y_{in_i}$ 之间是相关的, 这种潜在的相关性通常情况下是未知的, 其相关结构与问题的具体性质有关, 而且很难建模. 如何在纵向数据分析中对相关性进行建模, 这正是纵向数据分析研究的热点与难点.

在纵向数据研究中, 根据观测时间我们可以将纵向数据分为以下几种情况:

(i) 每个个体的观测时间相同, 并且相邻的观测时间是等时间间隔的, 即对任意的 i 和 $j(j \neq i)$, $t_{ik} = t_{jk}$, 且对任意的 i, $t_{i,k+1} - t_{ik}$ 是一个常数, $k = 1, 2, \cdots, n_i - 1$.

(ii) 每个个体的观测时间相同, 但相邻的观测时间不是等时间间隔的, 即对任意 i 和 $j(j \neq i)$, $t_{ik} = t_{jk}$, 但对任意的 i, 存在 $k \neq l$, 使得 $t_{i,k+1} - t_{ik} \neq t_{i,l+1} - t_{il}$.

(iii) 开始的测量时间可以不一样, 但时间间隔相同, 即 $t_{i,k+1} - t_{ik} = t_{j,k+1} - t_{jk}$.

(iv) 每个个体的测量时间可能不同, 相邻的观测时间也可能不是等时间间隔的, 即存在 i 和 j, $t_{ik} \neq t_{jk}$, 且对任意 i, 存在 $k \neq l$, 使得 $t_{i,k+1} - t_{ik} \neq t_{i,l+1} - t_{il}$.

对于上述四种不同情况, 其相关结构的假设可能是不同的. 例如对于 (i), 我们可以假设来自同一个个体的每两个相邻数据间的相关性相同, 且不同个体具有相同的相关结构; 而对于 (iii), 相关性可能依赖于时间间隔.

纵向数据在生物学、医学、社会经济学、心理学、教育学、环境科学等领域广泛存在. 近几年来, 纵向数据的统计分析引起广大统计学家和实际应用研究者的高度关注. 在经济学中, 纵向数据有另一个称呼, 称为面板数据 (panel data)(Cheng, 2005), 每一个个体称为一个面板. 在医学和生物学研究中, 还有一类数据称为集团数据 (clustered data), 它是指全部数据因为某种共性被划分为一些小的集团. 例如来自同一窝的小老鼠, 来自同一父母的子女. 因为相同的生活环境和相似的遗传基因, 来自同一个集团的数据间存在着某种相关性. 而纵向数据的相关性是因为数据在同一个个体上收集, 每一个集团可以被看作纵向数据中的一个被试, 故纵向数据统计分析的方法也可以用来分析集团数据. 但是集团数据不完全等同于纵向数据, 因为在纵向数据中, 来自同一个个体的数据因测量时间的不同而有先后顺序, 而集团数据则没有. 在后续章节中, 我们对纵向数据和集团数据不加区分.

1.2 纵向数据实例

在这一节, 我们介绍几个纵向数据的例子, 通过例子说明纵向数据研究中所关心的统计问题. 有些例子将会在后续章节中继续进行讨论.

1.2.1 HIV 数据集

CD4+ 细胞是人体免疫系统中的一种重要免疫细胞, 由于 HIV (艾滋病病毒) 攻击的对象是 CD4+ 细胞, 所以在医学研究中, CD4+ 的检测结果对艾滋病治疗效果以及对患者免疫功能的判断起着极其重要的作用. HIV 数据集收集了在多中心艾滋病队列研究 (Multicenter AIDS Cohort Study, MACS) 里注册的 369 名感染者自三年前到从血清转化之后的六年内的 CD4+ 细胞个数 (Kaslow et al., 1987; Diggle et al., 2002). 每个被试的 CD4+ 细胞个数被测量的次数从 1 次到 12 次不等, 一共有 2376 个数据可以利用. 协变量包含被试者血清转化时的年龄、吸烟状况 (以吸烟的包数来衡量)、是否使用药物、性伴侣的个数和用 CESD 量表测量的抑郁水平 (值越大表明抑郁状况越严重). 图 1–1 给出的是 20 名感染者 CD4+ 细胞个数随时间的变化情况, 图中号码是每名感染者的编号. 从图中可以看出, 每个被试的观测次数并不相等, 因此数据是非平衡数据. 随着时间的推

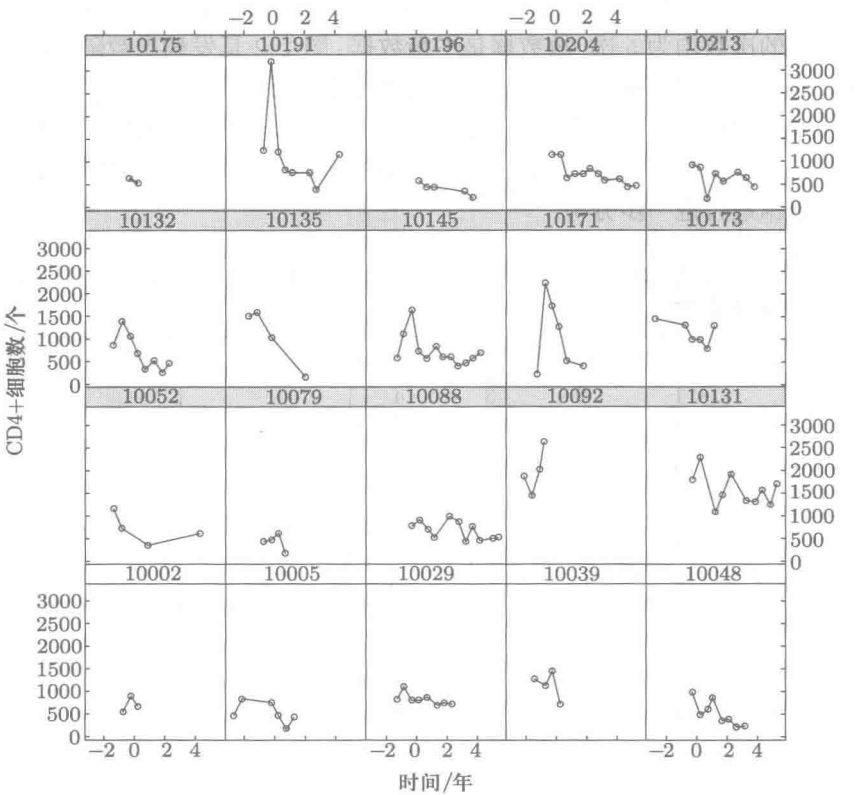

图 1–1 HIV 数据集中的 20 名感染者 CD4+ 细胞个数在血清转化前后随时间的变化情况

移, 大部分被试的 CD4+ 细胞个数都有下降的趋势. 此研究的主要目的是估计 CD4+ 细胞下降速度 (例如, 下降 50% 的平均时间) 以及这些协变量对 CD4+ 细胞数目的影响.

1.2.2 普罗加比药物研究

在一次临床试验研究中, 59 名癫痫症患者被随机地分为两组 (Leppik et al., 1985). 一组服用抗癫痫药普罗加比 (被试 1 ~ 28), 一组服用安慰剂 (被试 29 ~ 59). 研究的目的是判定抗癫痫药普罗加比能否降低癫痫症患者发病的次数. 因为每名患者的发病情况可能不同, 在开始进行分组治疗的前 8 周, 研究人员首先了解和记录了每名患者在这 8 周内癫痫发病的次数, 我们称其为基线水平 (baseline). 基线水平测量在临床研究中是很常见的. 在分组服药后, 每隔两周记录一次每名患者的癫痫发病次数, 共记录 8 周. 因为怀疑患者的治疗效果可能与其年龄有关, 因此对患者在接受治疗时的年龄也进行了记录.

表 1–1 给出了每组中的前 5 个被试在试验中的发病次数. 在该数据中, 每个患者观测的次数均为 5 次, 故数据是平衡数据. 图 1–2 是发病次数随时间变化趋势图. 该图表明, 在服用普罗加比组可能存在异常值. 两组在基线水平和服药后的平均发病次数似乎没有太大的差异. 从图 1–3 可以看出, 来自同一个被试的数据间存在着较强的相关性 (Thall & Vail, 1990). 我们将在后续的第四章和第六章中对该数据进行进一步分析.

表 1–1 普罗加比药物研究中的一组子数据: 安慰剂组 (0) 中的 5 个被试和普罗加比组 (1) 中 5 个被试的观测数据

患者编号	期间 (每两周)				处理效应	基线	年龄
	1	2	3	4			
1	5	3	3	3	0	11	31
2	3	5	3	3	0	11	30
3	2	4	0	5	0	6	25
4	4	4	1	4	0	8	26
5	7	18	9	21	0	66	22
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
29	11	14	9	8	1	76	18
30	8	7	9	4	1	38	32
31	0	4	3	0	1	19	20
32	3	6	1	3	1	10	30
33	2	6	7	4	1	19	18

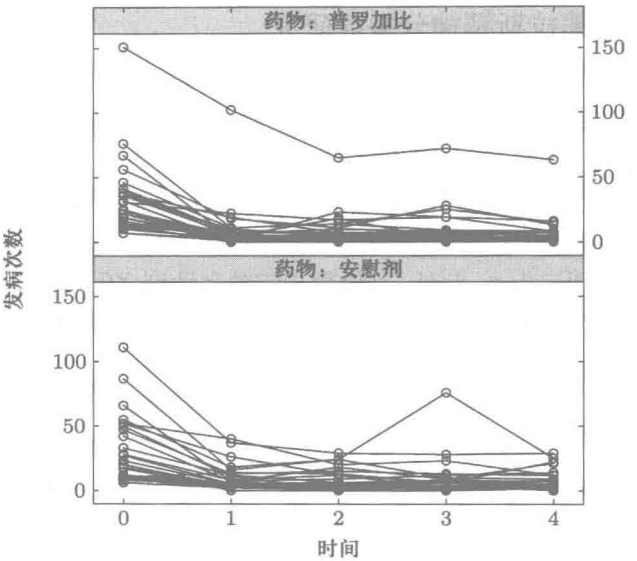

图 1–2 癫痫患者发病次数随时间变化趋势图, “0” 表示基线

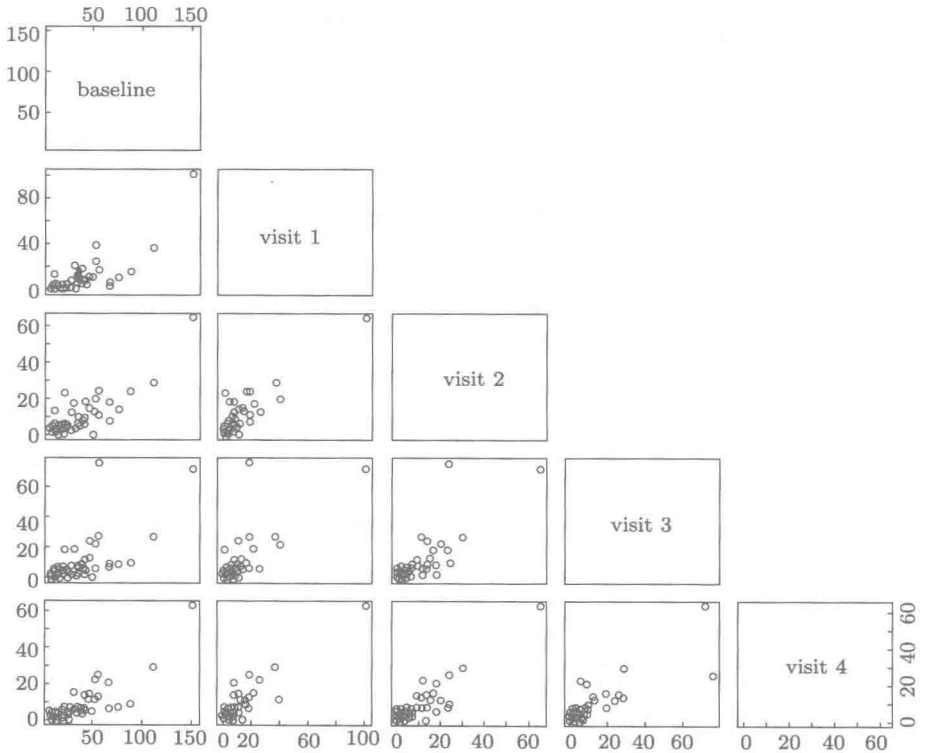

图 1–3 癫痫患者 5 次观测中发病次数的散点图

1.2.3 马德拉斯精神分裂症研究

本数据来自印度马德拉斯精神分裂症的纵向研究 (Heagerty & Zeger, 1998; Diggle et al., 2002). 该研究记录了 86 位精神分裂症患者在首次住院治疗后的第一年内是否出现思维障碍. 如果出现则记为 1, 否则记为 0. 因此, 该数据属于二分类数据. 协变量包括月份数 (患者首次入院治疗之后记录的月份); 患者发病时的年龄, 如果年龄不超过 20 岁, 则记为 1, 否则记为 0; 患者的性别, 如果是女性, 则记为 1, 否则记为 0. 此外, 月份数与年龄和性别之间都存在一定的交互效应. 该数据总共包含 921 次观测, 其中每位患者的观测次数大于等于 1, 小于等于 12. 该研究的主要目的是分析精神分裂患者在入院治疗后, 各个协变量是否会对患者的思维障碍产生影响. 表 1–2 给出了在每次观测中发病/未发病的不超过 20 岁/超过 20 岁的男性/女性发病人数. 可以看出, 随着时间的推移, 发病的人数越来越少. 记录结果表明, 在年龄小于 20 岁的发病患者中, 男性多于女性; 但是在年龄超过 20 岁的发病患者中, 女性要高于男性. 我们将在第六章对该数据进行详细分析.

表 1–2 精神分裂症研究中不同年龄段和不同性别的发病人数

年龄	性别	月份											
		0	1	2	3	4	5	6	7	8	9	10	11
		$Y=0$ (未发病)											
<20	男	12	9	12	13	18	17	20	20	23	24	24	23
	女	9	13	17	17	17	20	19	19	17	16	16	16
$\geqslant 20$	男	5	5	6	5	5	7	9	11	11	10	10	10
	女	4	7	5	8	10	12	15	15	14	14	14	14
		$Y=1$ (发病)											
<20	男	19	23	19	16	10	11	8	8	5	4	3	4
	女	15	10	6	5	4	0	1	0	1	2	2	1
$\geqslant 20$	男	7	7	6	7	7	5	3	1	0	1	1	1
	女	14	11	12	9	7	4	1	1	1	1	0	0

1.2.4 分娩阵痛研究

在该研究中, 83 名孕妇中的 40 名被随机地分到安慰剂组, 另外 43 名被分到了药物治疗组. 当宫颈扩张到 8 cm 时开始服药或者安慰剂. 孕妇要求在 180 min

内, 每隔 30 min 就自身疼痛的情况进行自我报告, 疼痛程度被记录在 100 mm 的尺子上 (0 表示不痛, 100 表示特别痛). 感知疼痛水平测量精确到 0.5 mm, 因此响应变量本质上是连续型随机变量. 从图 1–4 中可以看出, 随着时间的推移, 药物治疗组的疼痛程度逐渐减轻, 而安慰剂组的疼痛程度则逐渐增加. 在每个时间点, 安慰剂组的疼痛评分似乎都高于治疗组. 图 1–5 表明该数据不服从正态分布, 且概率分布是有偏的. 关于本研究的更多细节可以参考 Davis (1991), Jung (1996) 和 Wang & Zhu (2006).

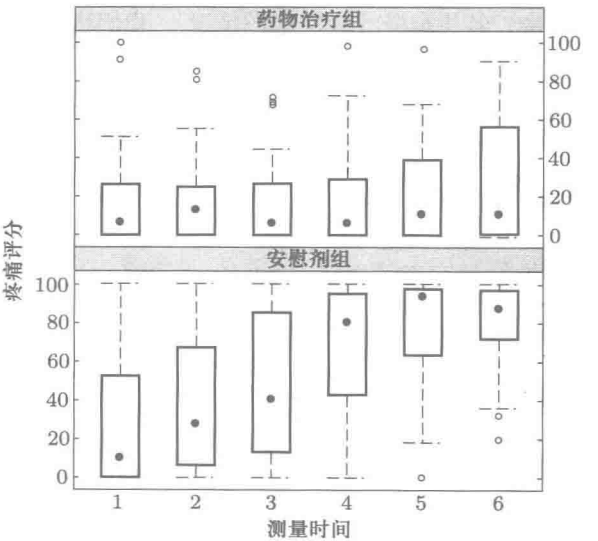

图 1–4 分娩阵痛研究中疼痛评分的盒形图

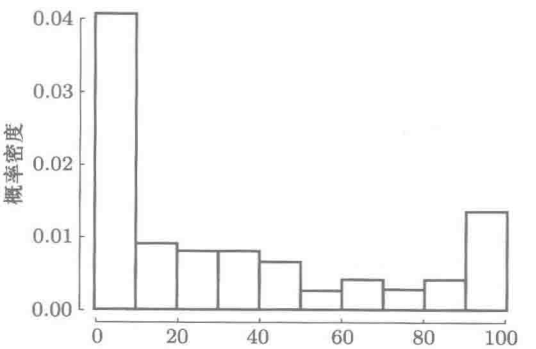

图 1–5 分娩阵痛研究中疼痛评分的概率密度直方图

1.2.5 呼吸道疾病研究

本数据是对美国俄亥俄州斯托本维尔市的 537 名 7 ~ 10 岁儿童是否患有呼吸疾病进行跟踪调查得到的. 如果调查时患病, 则记为 1, 否则记为 0. 该研究同时还调查了这些儿童的母亲在怀孕时是否吸烟 (吸烟记为 0, 否则记为 1)(Fitzmaurice & Laird, 1993). 本研究的主要目的是考察环境污染是否对儿童患呼吸道疾病有显著影响. 此外, 感兴趣的还有母亲吸烟和儿童患呼吸道疾病之间是否有关系, 以及感染呼吸道疾病的风险是否随着年龄的增长而发生变化. 表 1–3 列出了研究中 537 名儿童患呼吸疾病的情况. 我们将在第四章对该数据进行详细分析.

表 1–3 俄亥俄州斯托本维尔市 537 名儿童患呼吸疾病的情况

母亲怀孕时不吸烟					母亲怀孕时吸烟				
7 岁	8 岁	9 岁	10 岁		7 岁	8 岁	9 岁	10 岁	
			否	是				否	是
否	否	否	237	10	否	否	否	118	6
		是	15	4			是	8	2
	是	否	16	2		是	否	11	1
		是	7	3			是	6	4
是	否	否	24	3	是	否	否	7	3
		是	3	2			是	3	1
	是	否	6	2		是	否	4	2
		是	5	11			是	4	7

1.2.6 小老鼠病理试验研究

该试验是用来评估试验剂量对老鼠繁殖能力的影响. 在该研究中, 来自同一母体的幼鼠可以看作一个集团, 因此该数据属于集团数据. 在试验中, 30 只母鼠被随机地分到 3 个不同的处理组: 对照组、高剂量组和低剂量组, 每组均有 10 只母鼠. 在其产崽后, 分别测量每只幼鼠的体重. 但是, 在高剂量组中, 有一只母鼠不能生育, 另外一只母鼠吃掉了幼鼠, 还有一只母鼠生的是死胎, 最后仅有 27 窝幼鼠的数据可以用来分析. 幼鼠的体重被用来作为评估试验效应的指标. 图 1–6 给出了幼鼠的平均体重和每窝幼鼠只数之间的散点图. 图形表明, 药物的剂量越