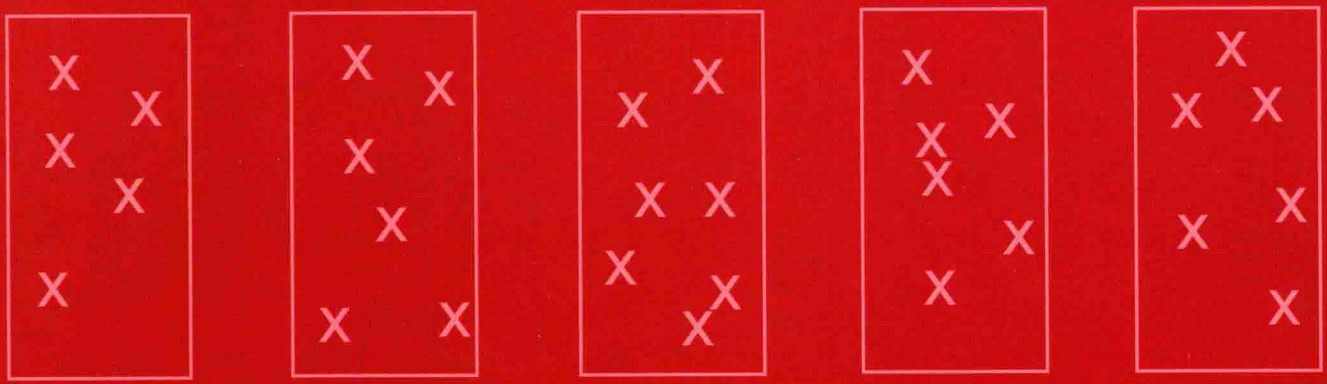
实用流行病学纵向数据分析方法

Applied Longitudinal Data Analysis for Epidemiology

第 2 版



原著 Jos W. R. Twisk

译者 陈心广
俞 斌
王培刚

RIDGE

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

实用流行病学纵向数据分析方法
Applied Longitudinal Data Analysis for Epidemiology

第 2 版

原著　Jos W. R. Twisk
译者　陈心广　俞　斌　王培刚

人民卫生出版社

图书在版编目(CIP)数据

实用流行病学纵向数据分析方法/(荷)乔斯·W. R. 特维斯克(Jos W. R. Twisk)原著;陈心广,俞斌,王培刚译. —北京:人民卫生出版社,2016

ISBN 978-7-117-23150-3

Ⅰ.①实… Ⅱ.①乔…②陈…③俞…④王… Ⅲ.①流行病学-数据处理-统计分析 Ⅳ.①R181.2-39

中国版本图书馆 CIP 数据核字(2016)第 203863 号

实用流行病学纵向数据分析方法

译　　者: 陈心广　俞　斌　王培刚
出版发行: 人民卫生出版社(中继线 010-59780011)
地　　址: 北京市朝阳区潘家园南里 19 号
邮　　编: 100021
E - mail: pmph @ pmph.com
购书热线: 010-59787592　010-59787584　010-65264830
印　　刷: 中国农业出版社印刷厂
经　　销: 新华书店
开　　本: 787×1092　1/16　**印张:** 17　**字数:** 392 千字
版　　次: 2016 年 9 月第 1 版　2016 年 9 月第 1 版第 1 次印刷
标准书号: ISBN 978-7-117-23150-3/R·23151
定　　价: 88.00 元
打击盗版举报电话: 010-59787491　E-mail: WQ @ pmph.com
(凡属印装质量问题请与本社市场营销中心联系退换)

作　者

Jos W. R. Twisk 教授，任职于荷兰阿姆斯特丹自由大学（Vrije Universiteit Amsterdam）医学中心、流行病与生物统计学系。Twisk 教授是应用纵向数据分析方法方面的专家，已出版关于应用纵向数据分析、多水平分析和应用生物统计概论等方面的书籍，并已发表论文 400 多篇。

译　者

陈心广教授，同济医科大学医学学士，卫生统计学硕士。夏威夷大学人口研究专业硕士，生物统计与流行病学博士。曾在美国南加州大学担任助理教授，韦恩州立大学担任副教授和终身教授，现任佛罗里达大学公共卫生学院流行病系终身教授，武汉大学珞珈学者讲座教授。

陈教授自从事科研工作以来，全面系统地开展了行为健康研究。在青少年艾滋病预防及控烟的行为干预研究方面，成功地将青少年危险性行为的干预项目移植到中低收入的发展中国家巴哈马，其部分研究成果被该国教育部和卫生部确定为本国中学生的“必修课”。在流动人口行为健康研究方面，开发了《个人社会资本评定量表》，并通过管理学、社会学、信息学等学科的高度交叉融合，成功将地球信息系统（GIS）、全球定位系统（GPS）技术和电脑辅助调查（ACASI）技术应用于流动人口行为健康研究中，2010 年在中国开展的流动人口艾滋病相关行为危险因素研究项目获得美国 NIH 资助。在基础理论与方法学研究方面，于 2008 年开创性提出了社会行为分析方法-概率离散事件系统模型（Probabilistic Discrete Event System Model），建立了从横断面调查数据提取纵向信息的方法，同年基于该理论的控烟项目获得美国 NIH 四年的资助。另外，陈心广教授成功将物理学中量子变化理论和混沌尖顶模型运用于行为健康研究领域，而且再次获得 NIH 五年的资助，深入开展研究，并于 2010 年 7 月受邀在美国混沌理论与心理学和生命科学学会年会上作主题发言。截止目前，陈教授共发表论文 200 多篇，在行为健康研究领域具有广泛影响。

俞斌，武汉大学公共卫生学院预防医学学士，武汉大学和美国韦恩州立大学联合培养流行病与卫生统计学硕士，毕业后任职于佛罗里达大学公共卫生学院流行病系，现在是该系的博士研究生。俞斌主要从事关于留学生文化同化和心理健康，农民工社会资本与健康行为，以及流行病方法学等方面的研究。

王培刚教授，法学博士，美国韦恩州立大学博士后，武汉大学全球健康研究中心教授，中组部青年拔尖人才，*Global Health Research and Policy* 编辑部主任，主要从事人口与健康、生活质量等领域的研究，获得国家社科基金、教育部人文社科基金等 10 多项课题资助。以第一作者和通讯作者的身份在国际 SSCI 期刊发表论文 10 多篇。

序

本书的读者对象为广大科研人员，包括高校和科研机构的研究人员和研究生。除了公共卫生和流行病学专业人士外，其他任何涉及纵向设计数据分析的科研人员，都可以参考使用本书。随着科学研究的深入发展，简单的横断面设计已经无法满足研究的需要，越来越多的科研课题开始采用纵向研究设计方案。相对于横断面设计，纵向研究的优势在于其可以帮助科研人员作因果联系推断。然而，由于纵向研究的设计特征，通常用来分析横断面设计数据的统计学方法大多数都不能有效地分析纵向数据。缺乏适当的统计学方法曾经一度限制了纵向研究的发展。

最近20多年来，新的统计分析方法（如GEE分析、混合效应模型等）不断出现，促进了纵向数据的统计分析。大量的统计学论文和专著从数理统计原理和算法方面，详细介绍了分析纵向数据的方法。为了促进这些方法的使用，本书作者避免了重复介绍复杂的数理统计学原理，强调了方法的具体操作和实际应用。为了更好地阐释这些方法，作者按照流行病学研究设计（即观察性研究和实验性研究）和结果变量的类型（即连续性、二分类、多分类和计数结果变量）系统地描述了纵向数据分析方法的应用。此外，作者还相应地介绍了传统的分析方法（如 t 检验、方差分析和卡方检验等）在不同类型数据中的应用，并列出了结果，便于读者作出比较。同时，本书作者还通过实际示例对每一种方法进行了介绍，并且列出了分析所采用的不同程序，如Stata、SAS、SPSS和R等，以及对结果的解释，极大地方便了读者的学习和应用。经过多番挑选和比较后，我们最终决定翻译本书，衷心地希望该书能够成为研究人员和院校学生的得力助手。读者可以通过网址（http://globalhealth.whu.edu.cn/Resource/Download/2016/0822/183.html）获取本书研究数据。如对本译著有任何批评性建议或相关疑问，请和俞斌（byu@phhp.ufl.edu或747939975@qq.com）联系。

由于本书中所涉及的统计学分析方法皆为最新，许多专业术语的中文名称尚未标准化。为了便于读者准确理解方法原理，并便于其阅读英文文献，文中多处附有英文原文，供读者参考。由于水平有限，翻译不当之处难免存在，请读者谅解。

本书由
武汉大学人文社会科学
青年学者学术发展计划
资助出版

前　言

本书最重要的特点在于其“应用”性，着重介绍了纵向数据统计分析方法的实际应用，而非“痴迷”于数理统计理论。在很多介绍纵向数据分析的书中，数理统计理论往往是其重要组成部分，这并不奇怪，因为几乎所有的此类书都是统计学家编写的。统计学家可以完全理解蕴含在纵向数据分析中的复杂的数学原理，但却很少有人能用简单的、能被研究人员所理解的方式来介绍这些复杂的数理统计原理，而研究人员往往更关注如何应用及怎样解释分析结果。本书则由一名流行病学家编写，很好地解决了上述问题，相对于统计学家，流行病学家更清楚如何运用恰当的统计分析方法，帮助寻找特定研究问题的答案，他们更多地关注如何应用统计方法，以及如何解释得到的结果。由于基本兴趣点和思考方式的不同，统计学家和流行病学家之间往往存在交流的障碍。除了对纵向研究日益增长的兴趣外，作者写这本书的初衷之一是帮助从事流行病学研究的人更好地使用统计学方法研究解决流行病学问题。除了流行病学外，本书也适合其他“非统计专业”研究人员。本书的目的旨在为处理纵向研究的数据提供实用性的指导，并为统计学家和流行病学家提供一个交流平台，共同探讨复杂纵向数据的分析问题。

目　录

第 1 章

纵向研究概论

1.1 背景知识

纵向研究（Longitudinal studies）指的是对结果变量进行两次及以上重复测定的研究，比如对同一个体的某一结果变量在不同的时间和条件下进行多次测定。由于纵向研究中每个个体在不同时间点所观测结果不是相互独立的，因此必须用专门的统计方法，以考虑每个个体重复观测的结果之间的相关性。必须指出，某些统计分析技术，如生存分析（Survival analysis）等，由于其不属于纵向数据分析技术的范畴，因此未纳入本书范围。生存分析所观察的结果变量（如死亡）通常是不可逆转的，虽然研究需要对个体进行追踪，但严格地来说，它只测量了一次结果变量。当所研究的事件发生之后，就不再继续进行观察。

为什么纵向研究在今天如此流行？原因之一就是研究者认为纵向研究可解决因果关系（causality）的推断问题。当然，这种理解也存在一定的偏差。表 1.1 列出了在每本流行病学书上都可以找到的有关因果关系判断的主要标准（e. g. Rothman and Greenland，1998）。其中只有一项适用于纵向研究：关联的时序标准（the rule of temporality），即在结果变量 Y（效应）和协变量 X（原因）之间有一个时间滞后，即原因必须发生在效应之前。因果关系是否存在，需要通过专门的纵向研究（例如，实验性研究）设计来判断，而并不是所有的纵向研究都可用来判断因果联系。那么纵向研究的优势是什么呢？一个纵向研究项目往往十分昂贵且耗时较长，同时数据分析也较为困难。如果和横断面研究（cross-sectional study）相比没有什么特别优势的话，为什么要费心费力去做纵向研究呢？通过仔细分析可

表 1.1　因果关系的判断标准

1. 关联的强度（Strength of the relationship）
2. 关联在不同人群不同条件下的一致性（Consistency in different populations and under different circumstances）
3. 关联的特异性：一种原因导致一种结果（Specificity：cause leads to a single effect）
4. 关联的时序性：原因在前结果在后（Temporality：cause precedes effect in time）
5. 生物学梯度：剂量-反应关系（Biological gradient：dose-response relationship）
6. 生物学上的合理性（Biological plausibility）
7. 实验证据支持（Experimental evidence）

以发现，相对于横断面研究，纵向研究的一个主要优势在于，它能够帮助我们观测结果变量随时间的独立变化过程。除此以外，结果变量的独立变化过程，还可以与其他变量的独立变化联系起来，从而分析影响结果变量的因素。

1.2 统计学的基本方法原则

本书介绍和解释统计分析技术的基本原则是“研究的问题决定所采用的分析方法”。尽管这一点似乎显而易见，但是意识到统计分析是为了解决特定的问题这一点是非常重要的。本书每章都以研究中碰到的具体例子开头，所以整本书会涉及很多在研究中碰到的实际问题。根据结果变量的特征，本书分为若干章节。每章都列举实际例子，以及研究所用的数据和统计分析的结果，以便于对统计分析的结果进行科学的解释。

1.3 分析纵向数据的知识基础

为了便于读者理解和掌握，本书力求通过深入浅出的方法，来介绍复杂的统计分析技术，同时将统计结果的解释与流行病学的研究问题相结合，但仍需要读者具备一定统计分析基础知识，如线性回归、Logistic 回归分析和方差分析等。

1.4 本书的示例

本书尽量用同一个纵向研究的数据库作为例子来介绍不同的统计分析方法。这个数据库由一个结果变量 Y 和 4 个协变量构成。结果变量 Y 是连续性变量，前后一共测量了 6 次。4 个协变量的分布特征不同（连续性或者二分类），且与时间的关联也不同，有时间独立的（Time-independent）也有时间不独立的（Time-dependent）。时间独立变量即变量的值不随时间改变而改变；时间不独立的变量即它的值随时间改变而发生改变。协变量 X_1 是连续的，也是时间独立的；协变量 X_2 是连续的，是时间不独立的；协变量 X_3 是二分类的，且时间不独立；而协变量 X_4 是二分类的，是时间独立的。所有时间不独立的变量和结果变量一样，都测量了 6 次。

在介绍结果变量为二分类变量的章节中（例如第 7 章），我们对连续性的结果变量 Y 进行了二分类化（Dichotomized）。而在结果变量是多分类变量的章节中（例如第 8 章），连续性结果变量 Y 被等分为三组。

本书所用的数据库，来自于阿姆斯特丹生长发育和健康研究项目。该研究是一个观察性的纵向研究，目的是为了分析青少年（Adolescents）和年轻人（Young adults）的生活方式与健康的纵向关系（Kemper，1995）。从单纯的统计分析角度来看，每个变量代表的具体含义并不是特别重要，

所以我们使用了较为简洁的符号来代表不同的变量。比如结果变量 Y 可以是任何变量，如一个特定的社会心理变量（Psychosocial variables）：抑郁量表的得分、生命质量的指标；或者是生物学参数（Biological parameter）：血压值及血液中白蛋白的浓度等等。在这个数据库中，结果变量 Y 实际上是指总血清胆固醇（mmol/L）；X_1 是在基线水平时的身体健康状况（Fitness level at baseline），以个体在跑步机上跑步时测得的最大氧气摄取量（Maximal oxygen uptake）来表示，由于是基线结果，因而与时间无关；X_2 是身体脂肪厚度（Body fatness），以身体四个部位的皮皱褶总厚度来判断，由于每次测量结果不同，因此与时间有关；X_3 是吸烟行为（二分类，吸烟或不吸烟，不同时候结果不同，因此与时间有关）；X_4 是性别（二分类，男或女，与时间无关）。表 1.2 是上述 5 个变量 6 次样本测量结果。

表 1.2　结果变量 Y 与协变量 X_1-X_4 的 6 次测量结果

时间点	Y	X_1	X_2	X_3	X_4
1	4.43（0.67）	1.98（0.22）	3.26（1.24）	143/4	69/78
2	4.32（0.67）	1.98（0.22）	3.36（1.34）	136/11	69/78
3	4.27（0.71）	1.98（0.22）	3.57（1.46）	124/23	69/78
4	4.17（0.70）	1.98（0.22）	3.76（1.50）	119/28	69/78
5	4.67（0.78）	1.98（0.22）	4.35（1.68）	99/48	69/78
6	5.12（0.92）	1.98（0.22）	4.16（1.61）	107/40	69/78

注：[a] 结果变量 Y 与协变量（X_1 和 X_2）是连续性变量，表中列出了均数和标准差；协变量（X_3 和 X_4）是二分类变量，表中列出了不同类别中个体的频数。

[b] 结果变量 Y 表示总血清胆固醇（mmol/L）。X_1 表示最大氧气摄取量［(dl/min) /$kg^{2/3}$］，X_2 表示皮褶厚度（cm），X_3 是吸烟行为（吸烟或不吸烟），X_4 为性别（男或女）。

[c] 本书中应用的数据库来自于 http://www.jostwisk.nl. 读者也可以通过网址（http://global-health.whu.edu.cn/Resource/Download/2016/0822/183.html）获取。

1.5　统计分析软件

在本书中，对那些相对简单的纵向研究的统计分析，我们使用了 SPSS 软件（version 18；SPSS，1997，1998.）。对比较复杂的纵向数据的分析，则采用了其他软件。对广义估计方程（Generalized estimating equation，GEE）和混合模型（mixed model）的分析，我们采用 Stata（version 11；Stata，2001），因为 Stata 的输出结果比较简洁明了，很适合分析复杂的纵向数据。在第 12 章，我们比较了不同统计软件在纵向数据分析上的应用特点，如 SAS（version 8；Little et al.，1991，1996）、R（version 2.13）、MLwiN（version 2.25；Goldstein et al.，1998；Rasbash et al.，1999）等。在所有的这些统计学软件中，复杂纵向数据的分析主要由软件完成。程序命令和结果输出也因不同的软件而有所不同。对于具体内容，可以参考相

关文献。

1.6 纵向研究的数据结构

在分析纵向数据时，不同软件对数据格式的要求也不同。纵向数据格式主要分为两种，“长”数据结构（“long”data structure）和“宽”数据结构（“broad”data structure）。在“长”数据结构中，每个变量多次测量的结果用一个变量表示，因此数据库中每个个体有多条记录，数据库从上到下会很长。而在“宽”数据结构中，每个个体只有一条记录，而同一个个体在不同时间和条件下测量的结果用不同的变量表示，数据库从左到右会很宽（图1.1）。

长数据结构

ID	Y	时间	X_4
1	3.5	1	1
1	3.7	2	1
1	3.9	3	1
1	3.0	4	1
1	3.2	5	1
1	3.2	6	1
2	4.1	1	1
2	4.1	2	1
.			
.			
N	5.0	5	2
N	4.7	6	2

宽数据结构

ID	Y_{t1}	Y_{t2}	Y_{t3}	Y_{t4}	Y_{t5}	Y_{t6}	X_4
1	3.5	3.7	3.9	3.0	3.2	3.2	1
2	4.1	4.1	4.2	4.6	3.9	3.9	1
3	3.8	3.5	3.5	3.4	2.9	2.9	2
4	3.8	3.9	3.8	3.8	3.7	3.7	1
.							
.							
N	4.0	4.6	4.7	4.3	4.7	5.0	2

图1.1 两种不同类型的数据结构

1.7 统计符号

本书中，我们尽量使用简单明了的统计符号，避免复杂的矩阵符号。个体的数量用 $i=1$ 到 N 表示，测量的次数用 $t=1$ 到 T 表示，协变量的数