万卷方法 | 社会科学研究方法经典译丛

Regression Analysis:
Statistical
Modeling
of a Response
Variable

附光盘

回归分析

因变量统计模型

鲁道夫 J.弗洛伊德 Rudolf J.Freund
威廉姆 J.威尔逊 William J. Wilson
平沙 Ping Sa 著
沈崇麟 译

重庆大学出版社
http://www.cqup.com.cn

万卷方法 | 社会科学研究方法经典译丛

■ 主编 沈崇麟 夏传玲

Regression Analysis:
Statistical
Modeling
of a Response
Variable

附光盘

回归分析

因变量统计模型

鲁道夫 J.弗洛伊德 Rudolf J.Freund
威廉姆 J.威尔逊 William J. Wilson
平沙 Ping Sa 著
沈崇麟 译

重庆大学出版社

Regression Analysis: Statistical Modeling of a Response Variable
by Rudolf J. Freund, William J. Wilson and Ping Sa

ISBN: 0120885972

版贸核渝字(2006)第 2 号。

图书在版编目(CIP)数据

回归分析：因变量统计模型/(美)弗洛伊德(Freund, R. J.)，(美)威尔逊(Wilson, W. J.)，(美)平沙(Sa, P.)著；沈崇麟译. —重庆：重庆大学出版社，2012. 9

(万卷方法)

书名原文：Regression Analysis: Statistical Modeling of a Response Variable

ISBN 978-7-5624-6976-6

Ⅰ. ①回… Ⅱ. ①鲁… ②威… ③平… ④沈… Ⅲ. ①回归分析 Ⅳ. ①O212. 1

中国版本图书馆 CIP 数据核字(2012)第 197691 号

回归分析：因变量统计模型

[美]鲁道夫 J. 弗洛伊德(Rudolf J. Freund)
[美]威廉姆 J. 威尔逊(William J. Wilson) 著
[美]平沙(Ping Sa)

沈崇麟 译

策划编辑：雷少波

责任编辑：李定群 高鸿宽 版式设计：雷少波
责任校对：刘 真 责任印制：赵 晟

*

重庆大学出版社出版发行
出版人：邓晓益
社址：重庆市沙坪坝区大学城西路 21 号
邮编：401331
电话：(023) 88617183 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
自贡兴华印务有限公司印刷

*

开本：787×1092 1/16 印张：25 字数：562 千
2012 年 9 月第 1 版 2012 年 9 月第 1 次印刷
印数：1—4 000
ISBN 978-7-5624-6976-6 定价：68. 00 元(含 1 光盘)

万卷方法学术委员会

总序
社会研究方法的现状及其发展趋势

近年来，社会调查技术和社会研究方法都有很大的发展。在调查技术方面，自20世纪70年代以来，社会变迁多次横断面的跟踪调查研究，几乎成为所有国家和地区了解社会结构转变和社会发展状况的基础性调查。这种调查不仅对社会学的研究有很大促进作用，而且对整个社会科学的研究都产生了重大影响，并且这些调查结果有的已作为政府有关部门决策的重要依据。国际上比较著名的此类调查有：美国芝加哥大学全国民意调查中心（National Opinion Research Center，简称NORC）的“社会综合调查（General Social Survey，简称GSS）”，英国埃塞克斯大学调查中心进行的“全国家庭生活和社会变迁调查”，法国经济和社会调查所进行的“全国经济社会调查”，日本社会学会组织进行的“全国社会分层与社会流动调查（简称SSM）”。中国台湾“中央”研究院社会学研究所，也每两年进行一次“台湾社会变迁基本调查”。美国的“社会基础调查”，现在已成为年度性的调查项目，它是美国国家基金会目前资助的最大的社会科学研究项目。以上这些调查，除美国的调查外，一般均因经费原因采用纵向的间隔性重复调查法，即每隔一段时间，进行一次全国规模的抽样调查。每次调查除保留社会研究所需的基本项目外，都有不同的主题。在间隔若干时间后，再重复同一主题的调查，这样的研究设计，使社会变迁研究在可以涉及更为广泛的研究领域的同时，具有更好的积累性和可比性。多年来，这些基础性调查获得的资料，滋养着大批的社会科学研究者，有时一项调查就有很多名博士生用来写博士论文，以此取得的研究成就，其可靠性受到社会科学界的广泛认同。例如1997年出版，以台湾地区社会变迁基本调查数据为基础的研究报告集《90年代的台湾社会，社会变迁基本调查研究系列二》收集论文16篇，内容涉及社会生活的各个方面，在台湾地区引起了极大的反响。

国内社会科学界在这方面也有了长足的发展。笔者所在的中国社会科学院社会学研究所的社会调查和方法研究室，组织或参与了多项与社会变迁有关的大规模抽样调查，取得了一定的研究成果，并积累了大量有关社会变迁的宝贵数据资料，其中主要有：

1. 城乡家庭变迁系列调查：该课题是由中国社会科学院社会学研究所牵头，联合北京大学和地方社科院的研究人员展开的一项类似多次横断面的城乡家庭变迁调查。这一调查始于1981年的“中国五城市婚姻家庭调查”，而后有1988年的“中国农村家庭调

查”、1991 年的“中国七城市家庭调查”、1998 年的“中国城乡家庭变迁调查”。

2. 有关中国城乡社会变迁的系列调查:这一调查始于 1991 年的第二批国情调查,然后有 1992 年的“中国城乡居民生活调查”、1993 年的“第三批国情调查”、1995 年的“第四批国情调查”和 1997 年的“中国沿海发达地区社会变迁调查”。上述调查虽然还不是严格意义上的多次横断面的纵贯研究,但研究者已在研究设计中尽量考虑到纵贯研究的基本原则,如调查队伍的稳定、指标的可比性和样本空间的延续性等。

3. 中国城乡社会变迁调查:这一调查始于 2000 年,为中国社会科学院重大课题。目前已经完成第一期第一次调查和第二次调查,今后将把这一调查发展为连续的、定期进行的社会变迁调查。

在纵向调查技术取得长足进步的同时,20 世纪末至今,电话调查也有很大发展。电话调查涉及的范围几乎与个别(面对面)访谈同样全面。电话调查中使用的一系列方法,是在 20 世纪 70 年代后期和面对面调查一起发展起来的。在 20 世纪 80 年代中期,电话调查开始变得很普遍,并且成为许多场合中各种调查方法的首选。正如某些学者所言,一种在公共和私营部门被人们用来帮助提高决策效率的收集信息的有效方法为人们所普遍认同时,这一现象本身就具有方法论上的意义。不仅如此,电话调查还有很大的实践意义,因为它为研究者提供了更多的控制调查质量的机会。这一机会包括抽样、被调查人的选择、问卷题项的提问、计算机辅助电话访谈(CATI)和数据录入。正因为如此,今天在各种社会调查中,如果没有发现其他重要的足以放弃使用电话调查的原因,电话调查由于其独特的对调查质量进行全面监控的优点,常常成为各种调查方式的首选。由笔者翻译,重庆大学出版社出版的《电话调查方法:抽样、选择和督导》一书,也于 2005 年面世。

无论是纵向调查抑或电话调查,实际上都是收集研究资料的方法,而应用社会科学的发展,不仅在于调查技术,即收集资料技术的发展,还在于研究方法和分析技术的发展。近年来,无论是定性研究方法,还是定量研究方法都有了长足的发展。

首先,计算机技术的发展可谓突飞猛进,它对当今社会生活的各个方面产生了巨大的影响,在悄悄地改变着社会科学的研究风格和研究方式的同时,也大大提升了社会科学学者的研究能力。这种影响表现在研究过程的各个阶段,从理论建构(概念映射)、问卷设计(专业的问卷设计软件)、调查实施(计算机辅助访谈、计算机辅助电话访问系统、网络在线调查系统)、数据录入(光学标记识别软件)到数据分析(包括文本、声音、图像资料的处理),甚至延伸到写作发表阶段。这样的过程发生在如社会学、经济学、政治学、心理学、教育学中,促进了学科之间的相互借鉴和交叉融合,至少在研究方法上呈现出这种趋势。随着计算机计算能力的大幅度提高,20 世纪 80 年代后期,统计学领域内发生了一场“革命”,主要表现在对定类和定序变量的建模能力的大幅度提高上,以及与分布无关的统计分析模型的发展之上,特别是基于“Resampling”(包括 Bootstrap、Jackknife、Monte Carlo 模拟等)的建模技术。同时,计算能力的提高还带动了基于神经网络、动态模拟、人工智能、生态进化等新兴的分析和预测模型的发展。这些进展都为定量社会科学研究提供了更多的可供选择的工具。

亚德瑞安·E. 拉夫特里(Adrian E. Raftery)依据社会学家所处理的数据类型,将定量社会学在美国的发展划分为三个时代:第一代起始于20 世纪40 年代,交互表是其主要处理对象,研究重点是关联度和对数线性模型;第二代起始于 20 世纪 60 年代,主要处理单层次的调查数据,Lisrel 类型的因果模型和事件史分析是其研究重点;第三代起始于 20

世纪80年代后期，开始处理诸如文本、空间、社会网络等非传统的数据类型，目前尚没有形成成熟的形态。拉夫特里的综述，虽然更强调定量社会学研究对统计学的贡献，但也大致勾勒出定量社会学在国外的发展脉络。

从分析模型的角度来看，定量分析在以下几个方向有了突破性发展：

1. 缺失值处理：由于社会生活的复杂性，社会调查数据常常出现缺失值，传统的处理方式是忽略这些缺失值，或者用均值替代。但现在则倾向于用多重插值法（multiple implation）或者其他基于模型的方法进行处理。这些技术的发展，不仅会增强我们对数据的处理能力，而且将改变我们设计问卷的方式。基于这些技术，我们在不增加被访者负担的前提下，大大增加了调查问卷的内容：每个被访者只回答问卷的一部分，然后通过对缺失值的处理，获得他们对未回答部分的估值。

2. 非线性关系：线性假定是经典定量分析的一个常见假定，但在实际研究当中，线性假定只能被看作是对社会现实的一个逼近和简化。面对具体的研究数据，如果没有理论上的明确指引（不幸的是，我们常常没有中程理论的指引），我们是无法在线性模型和非线性模型之间作出取舍的。但MARS模型的出现，让我们可以从经验数据当中获得最为拟合的变量之间的函数关系，而不必预先作出线性假定。这样，理论思考和数据分析就可以实现一个互动的循环过程，定量分析就不单单是对理论和假设的简单证伪过程，而是理论思维一个重要组成部分。

3. 测量层次：20世纪六七十年代的统计模型，大多要求数据的测量层次在定距以上，如因素分析，但社会学的调查数据却大多为定类或定序数据。对应分析、Loglinear、Logit、Logistic Regression、潜类分析、Ordinal Regression、Normal Ogive Regression等统计模型的出现，大大提高了定量社会学处理定类和定序数据的能力。

4. 测量模型：基于文化、社会、心理和认知等方面的考虑，在社会学界仍有人对问卷调查在中国的效度提出质疑。抛弃“本土化”的文化执著，我们更应当关注的是问卷调查的项目反应理论（item response theory），即被访者回答问卷题器时的过程模型。这方面的进展主要表现在两个方面：一是分解测量量表的成分，如Rasch model、IRT分析、Mokken分析等；二是将测量模型与因果模型或其他分析模型结合在一起，明确把测量误差引入到分析当中，充分评估它们对分析结果的影响，如结构方程模型。

5. 潜变量模型：与测量模型相关联的另外一个发展方向是潜变量模型，例如，潜变量分层分析（latent class analysis）、潜变量结构分析（latent structure analysis）、潜变量赋值分析（latent budget analysis）等。“潜变量”这一概念表明，我们可以通过测量“显变量”来测量无法直接观察的理论概念，如权力、声望、地位等。这样，理论和现实之间，通过“潜变量”到“显变量”的映射（测量过程），就有了连接的桥梁。

6. 分析单元的层序性：在定量分析当中，我们常常强调要避免出现“生态谬误”，即分析单元的层次和结论或推论的层次不一致。与其相关的方法论争论是“宏观和微观”的问题。随着多层次模型的出现，我们可以同时考察多个层次上的问题，我们可以把个人放在其家庭背景中，再把家庭放在社区的背景下，考察个人层次的变量对社区变量的效应，或者社区层次的变量对个体行为的具体影响。在定量分析模型当中，“宏观和微观”的连接获得了建模技术上的支持。在这个领域当中，还有一个方向也值得关注：分析宏观层次的数据，对微观层次进行推论。

7. 社会网络模型：区分“关系数据”和“属性数据”，是把分析重点从个体/群体等社会单元转移到这些社会单元之间关系的第一步，社会网络模型是目前发展较快的一个定量

分析领域,其理论根基是结构主义。社会网络分析目前仍然具有较浓厚的"形态学"特征(基于图论的缘故),但却为我们理解社会关系在社会空间上的形态奠定了基础,通过计算机模拟和研究社会网络的历期数据,研究社会结构的"发生学"性质模型也处在萌芽状态当中。

8. 系统动力学:如果说社会网络模型是在社会空间上拓展定量社会学的研究手段,那么社会过程在时间上和物理空间上的属性,则是事件史模型、事件数模型、历期分析、Cox 回归、时间序列分析、Cohort 分析、状态空间模型等模型的研究对象。在这个领域,计量经济学为定量社会学研究提供了许多有益的范例。

9. 预测模型:上述模型仍然是在分析主义的范式下。有些社会学的应用研究,更强调模型的预测精度,而不是模型的认知价值,例如,社会趋势的预测。由于计算能力的提高,神经网络、基因算法、人工智能、模式识别等数据挖掘技术有了长足发展,已经出现了许多拟合经验数据的预测模型,比较成功的应用出现在计量经济学领域(如对股市的预测)。

10. 计算机模拟:对于社会学应用研究而言,研究的对象具有历史性、规模大、变迁的过程不仅漫长且表现某种渐进性的特点,且因社会隔离/社会伦理原因无法接近或有实验禁忌等,无法直接进行观察和研究,这时计算机模拟就成为一个可供选择的替代方案。计算机模拟主要有两个类型:一是基于计算机网络的模拟:每台微机作为一个代理,整个网络作为"社会"实时演化,如法国的 Swarm 计划;二是基于概念模型的系统,在计算机时间上,按照既定规则运行,较有名的研究是罗马俱乐部的《增长的极限》,常见的软件有 Simul, Arena 等。自然科学家对此方向似乎比社会学家更有兴趣。

定性研究方法一直是社会学研究领域中比较传统的研究方法,在社会学研究的古典时期,它甚至是社会学家手中唯一的研究方法。但随着定量研究方法在社会学研究中的广泛应用,定性研究方法就似乎越来越不受人们的重视。但需要澄清的事实是,在定量分析模型取得飞速发展的同时,在过去的二十多年里,定性研究方法也有了长足的进步。主要表现在以下六个方面:

1. 研究素材日益扩大:除了传统的参与观察、深度访谈、专题小组访谈之外,会话、交谈、电视、广播、文档、日记、叙事、自传(autobiography)等社会过程中自然产生的素材,甚至社会学理论本身(理论的形式化),也开始进入定性分析的视野当中。所有这些资料,不仅可以以文本的格式存储,而且,新型的多媒体介质,如图像、声音和视频,作为原始的分析素材,也日益成为定性分析的新宠。

2. 分析方法更加多样:定性方法的种类在最近的二十多年中,更是有了一个质的飞跃。在比较传统的、源自语言学的方法,如内容分析、话语分析、修辞分析、语意分析、符号学、论据分析等方法之外,社会学家也创造出自己独特的定性分析方法,如施特劳斯(Strauss)等人的扎根理论、海斯(Heise)的事件结构分析、拉津(Ragin)的定性对比分析、Abbott 和 Hrycak 采用最优匹配技术的序列分析、亚贝儿(Abell)的形式叙事分析(formal narrative analysis)、鲍尔(Bauer)等人的语库建设、Attride-Stirling 等人的主题网络分析和神经网络技术应用的定性分析领域。所有这些方法的一个共同特征是,把定性研究向更加系统、更加精确、更加严格、更加形式化的方向推进。

3. 认识论基础更加多元化:现象学、释义学和本土方法论(ethnomethodology)的认识论,一直是定性分析的大本营,但近年来,实证主义也开始逐渐为定性分析所接纳,解释和阐释之间,由激烈的对立关系,逐渐演变为相互融洽的关系。

4. 研究过程更加客观规范：定性分析的一个主要问题在于阐释过程中不可避免的主观性。为了尽可能消除“解释者偏见”和主观选择性，定性分析开始遵循严格的程序模板或程序规则，并尝试引入定量分析中的“信度”“效度”“代表性”等概念，通过编码和对比，再加上传统的定性分析标准，如可解释性、透明性和一致性，使得定性研究的过程更加规范、阐释的结果更加客观，研究的结论更加可信。

5. 研究过程更加有效率：这主要应归功于大量计算机辅助定性数据分析（CAQDA）软件的涌现。从20世纪80年代以来，定性分析过程的数字化和计算机化，已经是一个不可逆转的大趋势。这种发展趋势与定性研究者的理论取向无关，不管他们的理论立场是实证主义、符号互动论，还是本土方法论，大多数定性研究者都在自己的研究当中，开始采用计算机来辅助定性资料的分析过程。据不完全统计，目前已经有二十多种定性分析的软件，分别隶属于德国、英国、法国、美国等国家。其中，有一些软件是国外研究机构的科研成果，可以免费使用，但比较成熟的定性辅助系统大多是商业软件。这些定性分析的辅助系统，不仅使得研究者从处理大量文字材料的繁复劳动中解放出来，而且能够让研究者共享他们各自分析的细节，从而改变定性研究的流程和研究集体之间的合作方式。同时，由于采用数据库结构，定性资料的管理也更加方便，这就为组织大型定性研究项目（包括多个研究地点、多个研究对象、历时的定性研究）提供了新的可能性。越来越多的定性研究人员开始走出他们的摇椅，坐到计算机屏幕前、湮没在访谈资料和故纸堆中的定性社会学家的形象已经一去不复返了。

6. 定性研究和定量研究的结合更加紧密：在定量分析方法的教材中，定性研究常常被看做是定量研究的前期准备工作，但定性研究者却持完全相反的观点，他们一般认为定性方法是自成一体的，可以完成从形成概念到检验假设的全部研究过程。在实际的应用研究中，定性方法和定量方法常常是交织在一起的，例如，克劳（Currall）等人在研究组织环境重要的群体过程时，通过内容分析把5年的参与观察资料量化，然后用统计分析来检验理论假定。格雷（Gray）和邓斯坦（Densten）在研究企业的控制能力时，利用潜变量模型把定性方法和定量方法有机结合在一起。雅各布斯（Jacobs）等人在研究比利时的家庭形态对配偶的家庭劳动分工影响时，首先用定量方法对纵向调查数据进行分析，从定量分析的结果中，又延伸出对核心概念的定性研究。这三个研究分别代表了定量和定性方法相互融合的三个方向：①克劳等人的研究代表着定性方法的实践者试图将定性数据尽可能量化的取向，近年来涌现出的处理调查数据中开放题器的编码问题的工具软件（如Words at，Smarttext等，注意：它们都是由著名的统计软件公司出品的处理定性资料的软件），处理定性资料的传统内容分析软件（如Nvivo、MaxQDA、Kwalitan等）也开始提供将定性资料转换到常用统计软件的数据接口，这些工具上的革新将加快这种趋势的发展。②格雷和邓斯坦的工作代表了“方法论多元论”的取向，即在应用研究过程中，通过核心概念的测量模型，把定性研究和定量研究结合在一起。③雅各布斯等人的工作则代表了一部分定量研究者对过度形式化的定量方法的不满，并试图通过定性方法加以弥补。在定量研究领域中，对“模型设定”问题的关注，是定量方法重新试图返回定性研究这种取向的另外一种表现。

与社会调查技术和社会研究方法突飞猛进的现实相比，我国学术界在这些方面的论著的出版似乎显得有些迟缓。虽然已经翻译了美国的一小部分经典定量分析教材，如布莱洛克（Blalock）和巴比（Babie）的教材，也有自己编写的一些教材，如袁方等人的《社会研究原理和方法》、卢淑华的《社会统计学》等，此外，偏重软件操作的还有郭志刚的《社会

统计分析方法——spss 软件应用》、郭志刚的《logistic 回归模型——方法与应用》、阮桂海的《spss for windows 高级应用教程》等。在《社会学研究》等专业杂志上,也常常有一些定量分析的应用研究,可是专门的方法和应用模型研究却没有,也没有专门的方法研究期刊。仅就定量研究方法的介绍而言,也存在一些缺陷,主要表现在:

1. 原理和操作脱节。
2. 过分依赖某些商业软件,不全面。
3. 与中国的实证研究相脱节。
4. 不能反映当前方法研究的最新进展。

与定量研究方法相比,由于各种原因,定性研究方法的引进和介绍都比较少。在福特基金会资助的方法高级研讨班上,曾讨论过一些定性研究方法。在定性方法研究方面也有少数专著,如袁方和王汉生 1997 年出版的教程,陈向明 2000 年出版的专著。但总体说来,我们对定性研究方法还停留在初步介绍的阶段,主要的介绍也局限在定性研究的研究设计和资料收集的阶段上,对定性分析方法的介绍,则没有能够反映出当代定性方法的最新进展。特别是在定性分析工具(定性分析软件)的引进和研究上,基本上还是一个空白。虽然不乏一些出色的定性研究报告,但从方法研究上讲,我们才刚刚起步。当然,我们同时还应该注意到,在历史学领域,我国对定性资料的鉴别、考据和分析,积累了大量的经验和知识,这也应当是定性方法研究的知识来源之一,应努力发扬光大。

令人欣慰的是,社会研究方法的引进和出版方面相对滞后的状况终于有所改观。重庆大学出版社的编辑,以独到的学术眼光,逆当前出版界唯利是图的不良选题风气,投入了大量的人力、物力,组织出版"万卷方法"。自 2004 年至今,已引进社会科学研究方法方面的专著十余种,在我国社会科学界已经引起了一定的反响。然而,更为可贵的是,重庆大学出版社并未以已经取得的成绩而自满,而是再接再厉,在原有"万卷方法"的基础上,进一步组织出版"万卷方法—社会科学研究方法经典译丛"。按我们的设想,"译丛"应该是一个开放的体系,旨在跟踪社会科学研究方法发展的前沿,引进和介绍这一方面的经典著作和最新成果。

"译丛"第一批有《抽样调查设计导论》《社会科学研究设计原理》《社会科学研究测量原理》《社会科学研究分析技术》《问卷设计手册》《回归分析》《数据再分析法》《抽样调查设计导论》《社会网络分析法》《广义潜变量模型》《分类数据分析》和《复杂调查设计和分析方法》(书名也许有变化)等十余种,几乎囊括了研究设计、测量和分析方法的所有领域,涵盖从基础的回归分析到最前沿的潜变量分析和多水平模型等各种分析方法。无论是社会科学各专业的本科生、研究生,还是社会科学研究的学者都将从中有所收获。

"译丛"由中国社会科学院社会学所社会调查与方法研究室的多位研究人员担纲,主译者都是在社会研究方法各个领域中具有相当造诣的教师和研究人员。"译丛"的译者不仅仅把翻译看作是一个"翻译",而且也把它看作是一次再学习和再创新。

我们期待"译丛"的出版能对社会研究方法的研究、应用和教学有所推动。

沈崇麟　夏传玲

2010 年 12 月于中国社科院社会学所社会调查与方法研究室

前　言

《回归分析:因变量统计模型》第2版旨在为用模型法对因变量做智能分析提供必需的工具。虽然本书的重点是介绍回归分析,但对其他线性模型,如方差分析、协方差分析和二分因变量的分析,以及非线性回归也有所涉猎。

我们普遍会遇到的问题是:手头上已经有了一组有关某一应变量的观察样本或实验数据,并希望通过统计分析对它的性状(behavior)做出解释。这种分析通常都基于变量的性状是可以为某一模型所解释的这样一个前提。而这样的模型(一般)的形式是涉及其他一些变量的代数表达式。那些其他的变量描述了实验条件、描述这些条件如何影响因变量的参数和误差。而误差表达式则几乎是无所不包的这一点,则说明任何模型都不可能对因变量的性状完全做出解释。统计分析包括参数估计、推论(假设检验和置信区间)和确定误差的性质(数量)。此外,我们还必须对那些有可能使统计分析出错的问题,如数据中的误差、模型选择不当和其他违反构成统计推论法的假设等进行调查。

用于这样的分析的数据既可以是实验、样本调查和过程的观察(操作数据)的数据,也可以是收集到的和第二手的数据。在使用所有这些不同来源的数据时,但尤其是在使用来自操作和第二手的数据做统计分析时,我们需要做的事不仅仅是将数目代入公式,或用一个计算机程序跑一跑数据。我们经常看到一些分析是由一些计划很差的一系列无序的步骤组成的。诸如这样的分析从定义、模型的构建、数据的筛选、计算机程序的选择,到输出结果的解释、数据异常之处和模型存在的不足的诊断,以及在分析目的的框架内提出的建言都可能存在着这样那样的问题。

注意,上面这些步骤中并不包括将数字代入公式。这是因为在分析过程中,这一工作是由计算机代劳的。因此,本书介绍的所有内容都假定计算工作是由计算机进行的,因而本书关注的只是做一个恰当的分析所涉及的其他方面问题。这就是说,本书将不会过多地讨论公式的问题,即使涉及公式问题,其目的也只是让大家了解计算机是如何进行分析的,当然偶尔也会对某些分析方法的原理做一些介绍。

为了使行文更有条理,本书将以如下的顺序来介绍本书涉及的各个专题:

1. 重温本书所需的基础课程。在简要介绍本书的内容和有关术语之后,再在线性模型背景下重温基本统计方法。

2. 全面复习简单线性模型。这一节的内容大多都是以公式为依据的,因为这些公式不仅都比较简单和有实际的解释,而且它所提供的原理对多元回归也很有意义。

3. 全面介绍多元回归问题,假设模型是正确的且数据是没有异常的。这一节也提到了几个公式,并使用了矩阵。对涉及的公式和矩阵,我们在本书的附录B中做了简要的介绍。不过这一节的重点是模型的构想、结果的解释、使用饱和及简约,或约束模型(full and reduced or restricted models)的参数推论,以及各种描述模型拟合情况的统计量之间

的关系。在结果解释时,特别关注*偏回归系数*(partial coefficients)的推导和结果的解释。

4. 介绍各种可以确定数据或模型出错的方法。深入浅出地介绍各种诊断潜在的问题的方法,以及各种可能对发现的问题的解决有所帮助的补救法。先从行诊断法(异常值和一些有关误差假设的问题)开始介绍,然后介绍列诊断法(多重共线性)。在对描述和推论两种统计工具进行介绍的同时,也对标准推论法在探索性分析中使用时应该注意的问题做了介绍。本章对变量的选择的方法步骤做了全面的介绍,但并非对它的一个方面等量齐观。我们更为关注的问题是,如何借助备择的变量选择法来对多重共线性问题进行补救。此外,本部分也讨论了行列问题之间的相互作用。

5. 介绍非线性模型。这一部分的内容包括那些可以用改造过的线性模型分析的模型,如多项式模型(polynomial models)、对数线性模型、二分自变量和因变量模型,以及严格的非线性模型和曲线拟合方法。曲线拟合法的讨论只限于拟合一条平滑曲线本身,与特定的模型无涉。

6. "广义线性模型"。这一模型既可用于连接方差分析(ANOVA)和回归分析,也可用于失衡的数据(unbalanced data)和协方差分析。

7. 用定类自变量分析定类因变量方法。

8. 用线性模型法分析非正态数据的自成体系的路数统称为广义线性模型(Generalized Linear Models)。这一部分涉及内容较本书其余部分的内容要更高深。因为所有的例子都是用 SAS 做的,所以《如何用 SAS 做线性模型》(SAS© *for Linear Models*, Littell et al, 2002)一书无疑是本书的最佳姐妹篇。

例子

对于一本讨论回归的著作而言,例子无疑是非常非常重要的。能称得上好例子的例子应该具备如下 3 个条件:

- 能为来自各个不同学科的学生所理解
- 含有数量适当的变量和观察
- 有某些令人感兴趣之处

本书列举的例子大部分都是"真的",因而通常都有某些令人感兴趣之处。为了易于理解和能令人感兴趣,我们对数据做了一些修改、删节或重新定义。有时我们也可能会编造一些例子的数据。我们也假设,在那些为某些特定的学生设计的课程中,教员或学生将会以课程专题的方式提供一些其他的例子。

为了保持行文的一致性,绝大多数例子都依据 SAS 系统的输出结果来讲述。书中有为数不多的例子的讲解,以其他系统的输出结果为依据。这样做的目的固然在于比较,但更重要的在于使读者明了,绝大多数计算机输出的结果提供的信息几乎是完全一致的。有时为了节省篇幅和避免混淆,我们也会对计算机输出的结果有所删节。不过,本书希望自己介绍的方法能在任何计算机软件上使用,所以本书所有有关计算机使用的讨论都是一般性的。那些专门的软件的使用方法的讲授则留给了本书的授课老师。

习题

做习题是学习统计方法的一个非常重要的部分。然而,由于计算机的使用,做习题的目的有了很大的变化。学生不必再亲手将数字代入公式,并保证计算得到的数字的精

确性,也不必再在做到了这两步之后,才去做下一个习题。而现在,计算的精确性基本上都是由计算机保证的,所以学习的重点也变成了如何挑选合适的计算机程序,以及如何用这些程序来得到希望的结果。不仅如此,这一变化也使得如何恰当地解释这些分析的结果,以确定是否还需要进行其他的分析这一问题变得很重要。总之,现在学生已经能有机会对结果进行研究,并对它们的用处进行讨论。因为学生的习题与如何适当地使用和解释分析结果有关,所以这就有可能会使学生花费相当多的时间去做习题,特别是第4章和第4章以后的那些开放性的习题,会令学生花费更多的时间。

因为恰当地使用计算机程序也是习题一个重要的组成部分,因此我们认为授课老师应该要求学生都动手做本书给出的例子。为此,我们把本书的例子所用的数据刻成了光盘,随本书一起发行。动手做这些习题给学生的不仅仅是更多的信心,而且能使他们得出与我们提出的结论有所不同的结论。

我们给出了一套精当的习题。许多习题,尤其是那些在较后的章节中给出的例子,一般都不存在所谓普遍正确的答案。正因为如此,方法和与之相关的计算机程序的选择变得十分重要。正因为如此,在每当谈到什么是恰当的分析时,我们只是有选择地给出有限的参考性的提示,有时甚至连这样的提示都不给。最后,我们希望授课老师和学生都会提供一些富有挑战性且令大家都感兴趣的例子。在各种统计课,如像本书介绍的回归分析的教学中,这样的做法是值得提倡的。

我们假设读者已经修过包括假设检验、使用正态、t、F 和 χ^2 分布的置信区间等内容的基础统计学课程。尽管本书的附录B已经对矩阵代数做了简要的介绍,但是如果读者有时间能专门修一下矩阵代数的基础课程,无疑将会对本书的学习有很大帮助。尽管我们并不要求学生专门去修微积分的课程,但是本书还是在附录C中,对用微积分进行最小平方估计的方法步骤做了简要的介绍。本书并未对最低需要掌握的计算机知识设限,但本书列举的绝大部分例子都是用SAS做的。因此,《SAS系统在回归分析中的应用》(**SAS©** *System for Regression*, Freund and Littell, 2000)一书或可被视为本书的姐妹篇。

本书封面上的照片是1986年发射的挑战者号航天飞船发射前所摄。那一次发射的失败是灾难性的,失败的原因是它的几个固体燃料火箭助推器中,有一个的连接处的O形环被烧穿了。在灾难发生之后,科学家和工程师对发射时的温度和O形环失灵之间的关系做了缜密的分析研究。他们所做的分析包括用概率比对数模型(logistic model),将失灵的概率作为发动时温度的函数来建模。分析的数据采自航天飞船以前的23次发射。本书第10和11两章对概率比对数模型做了详尽的介绍。飞船23次发射的数据,以及在SAS系统中用概率比对数回归所做的整个分析,可参见列特尔(Litell)等人的著作(Litell, et al., 2002)。

数据集

实际上,例子和习题的所有数据集都已经刻在本书附带的光盘上了(中文版将放在封底提供的资源网站上)。文件的名称可参见光盘中的文件《README》(说明)。

致谢

首先我要感谢我们的工作单位,德州农工大学(Texas A&M University)和北佛罗里达大学(University of North Florida)统计系,没有他们的合作和支持本书是不可能完成的。我也必须对本书的各位评阅者表达我的感激之情。他们的评阅使本书增色不少。

他们是:

- **帕特丽夏·布坎南**(Patricia Buchanan)教授,宾夕法尼亚州立大学统计系
- **罗伯特·高德**(Robert Gould)教授,加利福尼亚大学洛杉矶分校统计系
- **杰克·里弗斯**(Jack Reeves)教授,乔治亚大学统计系
- **詹姆斯·肖特**(James Schott)教授,中佛罗里达大学统计系教授
- **斯蒂文·格雷**(Steven Garren),詹姆斯麦迪逊大学数学和统计部
- **E. D. 麦昆**(E. D. McCune)教授,史蒂夫奥斯汀大学数学和统计系
- **K. 沙**(Arvind K. Shah)博士,南阿拉巴马大学数学和统计系

我们也要感谢 SAS 研究所,因为几乎所有例子的计算机输出结果都是用他们出品的软件(SAS 系统,the SAS System)来演示的。我们也用 SAS 系统制作了正态、t 和 χ_2 分布表。

最后,我们都要对我们的妻子表示深切的感谢之情。正是她们的鼓励,使我们在遇到挫折的时候没有半途而废,能坚持到了最后。

目 录