格致方法
定量研究系列

·革新研究理念
·丰富研究工具
·最权威、最前沿的定量研究方法指南

吴晓刚 / 主编
【美】威廉·D. 贝里(William D. Berry) / 等著

线性回归分析基础

格致出版社 上海人民出版社

格致方法
定量研究系列

吴晓刚 / 主编

【美】威廉·D. 贝里(William D. Berry) / 等著

线性回归分析基础

SAGE Publications ,Inc.

格致出版社 上海人民出版社

图书在版编目(CIP)数据

线性回归分析基础/(美)贝里(Berry, W. D.)等著;吴晓刚主编. —上海:格致出版社:上海人民出版社,2011

(格致方法·定量研究系列)

ISBN 978-7-5432-1898-7

Ⅰ. ①线… Ⅱ. ①贝… ②吴… Ⅲ. ①线性回归-回归分析 Ⅳ. ①O212.1

中国版本图书馆 CIP 数据核字(2010)第 254750 号

责任编辑 顾 悦

封面装帧 人马艺术工作室·储平

线性回归分析基础

[美]威廉·D. 贝里 等著

吴晓刚 主编

出 版 世纪出版集团 www.ewen.cc

格致出版社 www.hibooks.cn

上海人民出版社

(200001 上海福建中路193号24层)

编辑部热线 021-63914988

市场部热线 021-63914081

格致出版

发 行 世纪出版集团发行中心

印 刷 上海江杨印刷厂

开 本 787×1092 毫米 1/16

印 张 23

插 页 1

字 数 331,000

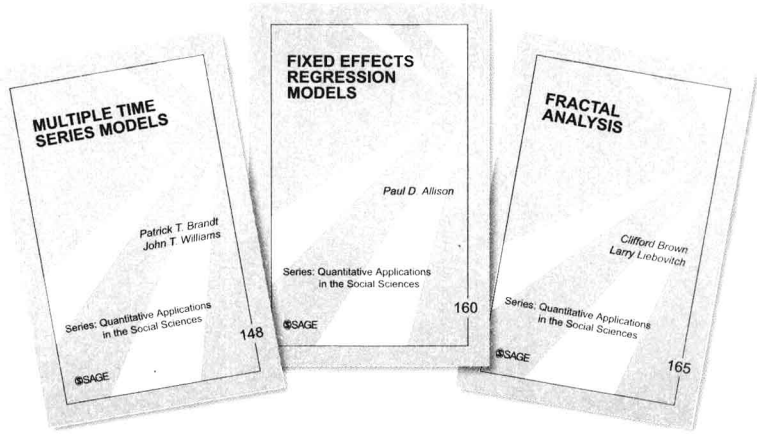
版 次 2011 年 7 月第 1 版

印 次 2011 年 7 月第 1 次印刷

ISBN 978-7-5432-1898-7/C·43

定 价 58.00 元

绿皮书的传奇



1976 年，当政府管理与政治学教授 Eric Uslaner 向 SAGE 出版社的创始人 Sara 建议出版关于重要定量研究方法的简明小册子时，没有人预见到这套书会取得巨大成功。

那年夏天，“社会科学定量研究方法” 丛书诞生。统一的、朴素的绿色封面，上面仅仅印着书名、作者名及本书所属的系列名。Iverson 和 Northporth 的《变量分析》(*Analysis of Variance*)、Nagel 的《运筹学》(*Operations Research*) 以及 Henkel 的《显著性检验》(*Tests of Significance*) 是最早出版的几本，每本售价 2.95 英镑，被形容为 “像烤饼一样好卖”。

SAGE 选择了 20 种顶尖的定量研究工具，然后去寻找合适的作者，邀请他们围绕这些工具撰写 92 页的小书。这些薄薄的绿皮书在课堂上深受欢迎，在图书馆成套陈列。

至今，绿皮书系列中共有 160 种在初版或再版。书的主题反映了量化研究方法的发展：从基础统计知识、数据类型、测量到计算机的应用以及博弈论。这套书非常畅销，其中最畅销的一本，是 1980 年出版的 Michael Lewis-Beck 的《回归方法的应用》(*Applied Regression*)。

出版说明

本书由四种讨论定量方法的小册子组成，分别是《理解回归假设》、《回归诊断简介》、《虚拟变量回归》以及《多元回归中的交互作用》。本书的主要内容如书名所示，是介绍社会学研究分析方法之一，即线性回归。线性回归分析是社会科学研究中最常见的分析方法，该书通过介绍回归分析的假设，接着质疑假设，进而提出新的变量分析方法，最后对回归分析中的各变量及其相互关系进行阐述，为读者提供了一套完整的对线性回归分析的认识。因此，该书的问世能向社会科学研究者提供更深入的理论指导。

《理解回归假设》能使研究者深入了解多元回归分析的假设，并更熟练地驾驭回归分析，完成更有效的估计。《回归诊断简介》针对回归中经常出现的影响估计精度的因素，对研究者的假设提出质疑，运用“回归诊断”判断假设的合理性并处理回归分析中存在的问题。《虚拟变量回归》针对回归分析中，定序或名义变量无法有效反映因变量与自变量之间的实际关系，提出“虚拟变量”之概念，完善回归分析。最后，《多元回归中的交互作用》对最小二乘法中存在的交互作用项进行分析，厘清模型中各变量之间的关系和互相影响的情况，并提出了许多新的问题（如聚类数据的交互作用等）。

总 序

往事如烟，光阴如梭。转眼间，出国已然十年有余。1996 年赴美留学，最初选择的主攻方向是比较历史社会学，研究的兴趣是中国的制度变迁问题。以我以前在国内所受的学术训练，基本是看不上定量研究的。一方面，我们倾向于研究大问题，不喜欢纠缠于细枝末节。国内一位老师的话给我的印象很深，大致是说：如果你看到一堵墙就要倒了，还用得着纠缠于那堵墙的倾斜角度究竟是几度吗？所以，很多研究都是大而化之，只要说得通即可。另一方面，国内（十年前）的统计教学，总的来说与社会研究中的实际问题是相脱节的。结果是，很多原先对定量研究感兴趣的学生在学完统计之后，依旧无从下手，逐渐失去了对定量研究的兴趣。

我所就读的美国加州大学洛杉矶分校社会学系，在定量研究方面有着系统的博士训练课程。不论研究兴趣是定量还是定性的，所有的研究生第一年的头两个学期必须修两门中级统计课，最后一个学期的系列课程则是简单介绍线性回归以外的其他统计方法，是选修课。希望进一步学习定量研究方法的可以在第二年修读另外一个三学期的系列课程，其中头两门课叫“调查数据分析”，第三门叫“研究设计”。除此以外，还有如“定类数据分析”、“人口学方法与技术”、“事件史分析”、“多层线性模型”等专门课程供学生选修。该学校的统计系、心理系、教育系、经济系也有一批蜚声国际的学者，提供不同的、更加专业化的课程供学生选修。2001 年完成博士学业之后，我又受安德鲁·梅隆基金会资助，在世界定量社会科学研究的重镇密歇根大学从事两年的博士后研究，其间旁听谢宇教授为博士生讲授的统计课程，并参与该校社会研究院（Institute for Social Research）定量社会研究方法项目的一些讨论会，受益良多。

2003年，我赴港工作，在香港科技大学社会科学部，教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课“社会科学里的统计学”(Statistics for Social Science)为所有研究型硕士生和博士生的必修课，而第二门课“社会科学中的定量分析”为博士生的必修课(事实上，大部分硕士生在修完第一门课后都会继续选修第二门课)。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的“消费者”而非“生产者”。作为“消费者”，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少量重复，但各有侧重。“社会科学里的统计学”(Statistics for Social Science)从介绍最基本的社会研究方法论和统计学原理开始，到多元线性回归模型结束，内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型，以及线性回归模型的假设和模型诊断。“社会科学中的定量分析”则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法，将重点放在因变量为定类数据的分析模型上，包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型，包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生，一直鼓励和支持我将两门课的讲稿结集出版，并帮助我将原来的英文课程讲稿译成了中文。但是，由于种种原因，这两本书拖了四年多还没有完成。世界著名的出版社 SAGE 的“定量社会科学研究”丛书闻名遐迩，每本书都写得通俗易懂。中山大学马骏教授向格致出版社何元龙社长推荐了这套书，当格致出版社向我提出从这套丛书中精选

一批翻译，以飨中文读者时，我非常支持这个想法，因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种语言的精准把握能力，还要有对实质内容有较深的理解能力，而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容，只有语言能力是远远不能胜任的。在短短的一年时间里，我们组织了来自中国内地及港台地区的二十几位研究生参与了这项工程，他们目前大部分是香港科技大学的硕士和博士研究生，受过严格的社会科学统计方法的训练，也有来自美国等地对定量研究感兴趣的博士研究生。他们是：

香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛，硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊，应用社会经济研究中心研究员李俊秀；香港大学教育学院博士研究生洪岩璧；北京大学社会学系博士研究生李丁、赵亮员；中国人民大学人口学系讲师巫锡炜；中国台湾“中央”研究院社会学所助理研究员林宗弘；南京师范大学心理学系副教授陈陈；美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛；美国加州大学洛杉矶分校社会学系博士研究生宋曦。

关于每一位译者的学术背景，书中相关部分都有简单的介绍。尽管每本书因本身内容和译者的行文风格有所差异，校对也未免挂一漏万，术语的标准译法方面还有很大的改进空间，但所有的参与者都做了最大的努力，在繁忙的学习和研究之余，在不到一年的时间内，完成了三十五本书、超过百万字的翻译任务。李骏、叶华、张卓妮、贺光烨、宋曦、於嘉、郑冰岛和林宗弘除了承担自己的翻译任务之外，还在初稿校对方面付出了大量的劳动。香港科技大学霍英东南沙研究院的工作人员曾东林，协助我通读了全稿，在此我也致以诚挚的谢意。有些作者，如香港科技大学黄善国教授、美国约翰·霍普金斯大学郝令昕教授，也参与了审校工作。

由于所选每本书都有一篇序言，对相关方法的背景和应用作了很好的介绍，我们均予以保留，内容在此不再赘述。为了方便起见，我们将内容相似的书目集册出版，每册三至五本不等，共八册，它们分别是：《线性回归分析基础》、

《高级回归分析》、《广义线性模型》、《列表数据分析》、《纵贯数据分析》、《因果关系模型》、《社会科学中的数理基础及应用》和《数据分析方法五种》。所冠书名未必能精准涵盖其中的内容，读者可自行参阅每本书的序言或目录。

我们希望本丛书的出版，能为推动国内社会科学定量研究的扎实学风作出一点贡献。

吴晓刚

于香港九龙清水湾

目 录

理解回归假设

序 3

第 1 章 简介 4

第 2 章 回归假设的正式描述 6

第 3 章 “体重”的案例 16

第 4 章 如何得到满意的回归假设结果 21

第 5 章 回归假设的实质意义 25

第 6 章 结论 79

注释 80

参考文献 85

译名对照表 87

回归诊断简介

序 91

第 1 章 概论 92

第 2 章 最小二乘回归 95

第 3 章 共线性 99

第 4 章　奇异值与强影响数据　109
第 5 章　非正态分布误差　126
第 6 章　不一致的误差方差　134
第 7 章　非线性　139
第 8 章　离散数据　145
第 9 章　最大似然法、计分检验和构造变量　150
第 10 章　建议　158
附录　163
参考文献　172
译名对照表　175

虚拟变量回归

序　179
第 1 章　简介　181
第 2 章　构建虚拟变量　187
第 3 章　虚拟变量回归　198

第 4 章　估计组影响差异　208
第 5 章　可替代虚拟变量编码方案　238
第 6 章　虚拟变量用法专题　249
第 7 章　结论　257
注释　258
参考文献　260
译名对照表　263

多元回归中的交互作用

序　267
第 1 章　导论　269
第 2 章　双向交互作用　283
第 3 章　三向交互作用　308
第 4 章　其他重要问题　323
注释　349
参考文献　350
译名对照表　353

理解回归假设

作 者 简 介

威廉·D. 贝里(William D. Berry)

曾于美国佛罗里达州立大学和肯塔基大学讲授统计学和研究方法,现为佛罗里达州立大学政治科学系教授。其主要研究领域是公共政策和美国政策,在学术期刊上发表了大量论文,还参与撰写了《理解美国政府的成长:对战后时期的经验研究》(Praeger, 1987)以及《实用多元回归》(Sage, 1985),同时他也是《非递归因果模型》(*Nonrecursive Causal Models*)(Sage, 1984)一书的作者。

译 者 简 介

余珊珊

2009 年毕业于清华大学社会学系,目前是香港科技大学社会科学部硕士研究生。

序

回归分析是社会科学研究中最基本的工具，至少对于非经验主义者而言是这样。尽管它是一件最常用的工具，但它同样有可能是最容易被滥用的工具。每位一年级的研究生都会快速地学习构造最基本的多元回归模型，比如：

$$Y = b_0 + b_1 X_1 + b_2 X_2 + e$$

我们假设政治学家 Betty Brown 利用如下最小二乘估计模型(OLS)估计美国 50 个州的福利花费情况：

$$\hat{Y} = 543.66 + 87.10 X_1 + 450.39 X_2$$

其中 $\hat{Y}$ = 各州的福利花费(美元/人)，X_1 = 民主党在国会里的议席(百分比)，X_2 = 城市人口(百分比)。

Brown 教授可能会总结道，民主党的议席每增加 1%，福利花费的期望值就会增加87.1 美元(当城市化水平保持恒定时)。那么这个对 X_1 效果的估计到底有多好呢？更确切地说，这是最好的线性无偏估计(BLUE)吗？如果答案是肯定的，那么这一估计模型就能够与真实的世界联系起来。否则，这一估计模型只是那些流连在铅笔和草稿纸上的平面。

显然，我们应该去寻找能够达到最佳无偏估计标准的估计模型。这是我们学习回归假设的原因。Berry 教授非常严谨地定义了每一个假设，并且阐述了它们的实质意义。这种优美的文字描述搭配精选的图形和通俗易懂的证明，使得那些难懂的问题，比如测量、设定、多重共线性、异方差性以及自相关，都变得平易近人。而本书中的案例和数据也安排得很有条理，模型中的一个变量更能广泛地吸引人们的兴趣——体重。

理解回归假设可以让研究人员看到自己的弱点，同时也能够使他们更好地驾驭回归分析，以得到更有效的估计。当然，没有这种理解，就无法迈开通往构建模型的步伐。尽管目前已经有很多著作涉及回归分析这一话题，例如《应用回归》、《回归分析的解释和应用》、《实用多元回归》、《随机参数回归模型》、《理解回归分析》、《多元回归中的交互影响》、《回归诊断简介》，但是还没有人专门研究回归假设。Berry 教授的贡献恰好能填补这一空白。

迈克尔·S. 刘易斯-贝克(Michael S. Lewis-Beck)

第 1 章 | 简介

在任何回归分析被运用到社会科学研究中的时候，研究者总是会或明确或含蓄地提出无数的假设。[1]社会科学的定量研究已经非常流行，以至于几乎所有研究生二年级的学生都能够背诵一长串标准回归假设。然而尽管学生们经常死记硬背这些假设，却不能够理解其中“真正的含义”。多年来，我常常与研究生们针对他们的研究交换意见。而下文中所出现的屡见不鲜的场景正是让我决定撰写本书的原因：

教授：在你的模型中，你对异方差性这个概念还有问题吗（或者对任何其他的概念——设定残差、测量误差、自相关、非线性等等）？

学生：我不知道。

教授：那么，异方差性指的是什么？

学生（自信地）：误差项的变化不是恒定的。

教授：好的。你的因变量是个人在慈善事业上的支出（或者任何其他变量）。你考虑了以下的自变量……在你的案例里面，如何解释误差项是异方差的？

学生（有点不自信了）：对于不同的观测值，误差项的变异会有不同的取值。

教授：告诉我，这对于你的模型而言实质上意味着什么？你怎么解释慈善支出、你的模型中的自变量、其他影响慈善支出但没有包含在你的模型中的因素，以及所有这些变量是如何联系起来的？

学生（意识到自己知识上的一些漏洞被发现了）：我真的不知道。

因此，尽管很多社会科学研究者能够自信地“不费吹灰之力地快速说出”一长串多元回归分析的假设（没有设定残差，没有测量误差，缺乏自相关等等），又或许他们能够说出这些回归假设的标准定义，但是常常缺乏对这些假设实质含义的深刻理解。如果我们对这些假设的理解仅仅局限于对定义的死记硬背，我们就无法把这些假设运用到对具体问题的分析中，这就相当于我们根本没有完全理解这些假设。

写作这本专题论著的目的是描述回归假设，并在某种程度上鼓励学生从死记硬背中解脱出来，转而去理解如何考察假设是否能够与一个具体的研究相适应。我们的讨论仅限于回归方法，因为回归在社会科学方法论中占据了主导地位，尽管也可以写出类似的关于其他的经验研究技术的著作。如果社会科学研究者能够仔细地考虑回归假设是否真正符合实际应用中的案例，而不是遇见什么问题都用回归方法来解决，那么当运用其他研究技术的时候，他们就能够更加自如地把握。

我以对标准多元回归假设的回顾作为开头，因为这些知识通常会出现在计量经济学或者回归分析的课本中。[2] 如果你不能理解这些假设的意义和重要性，不要担心。接下来，我会引入一个贯穿本书的具体案例，具体而言，这是一个关于体重的决定因素的模型。我选取这个案例是因为这里所涉及的人体的体重是与我们所有人都有关的话题——如果不考虑我们各自的兴趣——因此我们对此会有合理的直觉。最后，我回到回归假设，考察每一个假设的实际意义，并强调研究者如何评估每一个假设是否符合实际研究的需要。