格致方法·定量研究系列　吴晓刚　主编

现代稳健回归方法

[加] 罗伯特·安德森（Robert Andersen）著
李丁译

- 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

6

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

现代稳健回归方法

[加] 罗伯特·安德森 (Robert Andersen) 著
李　丁 译

SAGE Publications ,Inc.

格致出版社　上海人民出版社

图书在版编目(CIP)数据

现代稳健回归方法/(加)安德森(Andersen, R.)著;李丁译. —上海:格致出版社:上海人民出版社,2012

(格致方法·定量研究系列)

ISBN 978-7-5432-2141-3

Ⅰ. ①现… Ⅱ. ①安… ②李… Ⅲ. ①回归分析-研究 Ⅳ. ①O212.1

中国版本图书馆 CIP 数据核字(2012)第 157499 号

责任编辑　王亚丽

格致方法·定量研究系列

现代稳健回归方法

[加]罗伯特·安德森 著

李丁 译

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

发　行　世纪出版集团发行中心

印　刷　浙江临安曙光印务有限公司

开　本　920×1168 毫米　1/32

印　张　5.5

字　数　105,000

版　次　2012 年 8 月第 1 版

印　次　2012 年 8 月第 1 次印刷

ISBN 978-7-5432-2141-3/C·83

定　价　15.00 元

出版说明

由香港科技大学社会科学部吴晓刚教授主编的“格致方法·定量研究系列”丛书，精选了世界著名的SAGE出版社定量社会科学研究丛书中的35种，翻译成中文，集结成八册，于2011年出版。这八册书分别是：《线性回归分析基础》、《高级回归分析》、《广义线性模型》、《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》、《数据分析方法五种》和《列表数据分析》。这套丛书自出版以来，受到广大读者特别是年轻一代社会科学工作者的欢迎，他们针对丛书的内容和翻译都提出了很多中肯的建议。我们对此表示衷心的感谢。

基于读者的热烈反馈，同时也为了向广大读者提供更多的方便和选择，我们将该丛书以单行本的形式再次出版发行。在此过程中，主编和译者对已出版的书做了必要的修订和校正，还新增加了两个品种。此外，曾东林、许多多、范新光、李忠路协助主编参加了校订。今后我们将继续与SAGE出版社合作，陆续推出新的品种。我们希望本丛书单行本的出版能为推动国内社会科学定量研究的教学和研究作出一点贡献。

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。

我们希望本丛书的出版，能为建设国内社会科学定量研究的扎实学风作出一点贡献。

吴晓刚

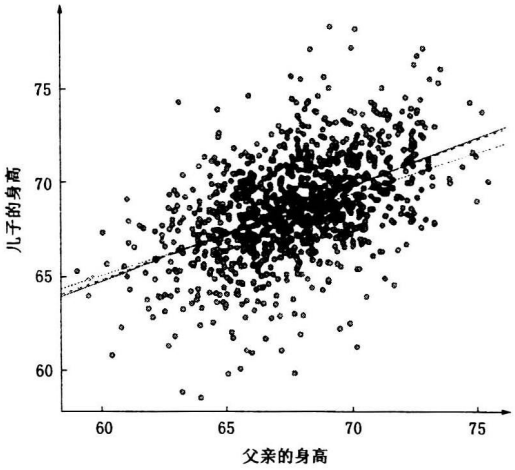
于香港九龙清水湾

序

1886年，弗朗西斯·高尔顿(Francis Galton)发表了题为《遗传身高向普通回归》(*Regression Towads Mediocrity in Hereditary Stature*)的开创性文章，从而开启了今天我们所知的线性回归统计方法的发展历程。通过分析205对父母及928个小孩的数据，高尔顿发现，相对较高或较矮的父母生养的小孩倾向于不是那么高或那么矮，这一特征被统计术语概括为“向均值回归”。

为了演示回归是如何处理此类身高数据的，我使用了一套相似但只有一个性别的数据，这应归功于高尔顿的学生卡尔·皮尔森(Karl Pearson)。下图标绘出了1078对父子的身高状况(单位是英寸)，数据用小圈点表示，它们明显地遵循一种线性趋势，刻画出向均值(等于45英寸)回归的现象。

在本图中，我拟合了一条回归直线，用实线表示，斜率估计值为0.514，由一般最小二乘估计得到(这一估计及以后其他估计的双尾检验都比常规的0.001水平显著得多，因此这里就不报告了)。不管以谁的标准来看，这一数据的表现都很不错。不过，即使是在这一表现良好的数据里面，有些案



例也比其他的更异常：我们很快就可看到图中右上角及左下区的某些案例离其他围绕在直线周边的大多数案例更远。如果这些案例太过极端，我们就可以从下列标准的快速“处理办法”中选择一个：从分析中剔除这些案例、重新编码（如果存在编码错误的话），以及在分析中纳入更多新变量。但如果没有处理这些异常（或不那么异常）案例的合理可用的解决办法，数据分析者该怎么办呢？这正是稳健及耐抗性回归方法（robust and resistant regression method）派得上用场的地方。

为了展示稳健回归，我对上述数据拟合了另外两条直线（使用的是R软件里的MASS数据包），虚线表示的是用*MM*估计量（*MM*-estimator）估计得到的稳健回归线（斜率估计值 = 0.502），点线表示的是通过将分位残差平方最小化（minimization of quantile squared residuals）的耐抗性回归估计（估计过程中分位残差最大的案例被忽略）得到的直线（斜

率 = 0.442)。可以看到,使用 *MM* 估计量得到的稳健回归结果,其斜率只比 OLS 回归的稍小。不过,耐抗性回归得到的估计结果差别更大,所得出的结论表现出更严重的向均值的回归。由安德森撰写的这本著作的焦点在于有效性(validity)的(而非效率的)稳健,它将帮助社会科学研究者理解这些方法,并学到稳健回归的原理及应用方法。

在社会科学中,现代稳健及耐抗性回归方法还不太为人所知。这些方法之所以被称为"现代方法",是因为它们通常属于密集型计算(computation intensive),这是当前很多依赖今天的高速电脑的统计方法的一个特征。本书(尤其是其中关于回归方法的那些章节)在主要统计软件如 SAS 和 Stata 已经采用这些最新回归方法的情况下是非常及时的。本书通过一套统一的符号系统介绍了不同来源的多种稳健回归方法以及它们彼此之间的联系,这正是本书的杰出贡献之一。为了给读者们一些实际应用上的帮助,本书也讨论了不同方法的相对优势和不足。通过一本这样的书,社会科学专业的学生及研究者最终会发现这些新的回归方法和经典回归方法一样平常和容易使用。

廖福挺

目 录

第1章

导　论

在定量社会科学中，回归分析是统计方法的主要干将。大量的问题都是由线性模型或者广义线性模型(generalized linear model)解决的。只要被恰当使用，回归估计就能为数据里的关系提供有效而简洁的概括。但如果盲目而机械地使用，回归分析也会导致错误的结论。异常观察案例的存在就是引起担心的原因之一，它们有时足以严重扭曲由一般最小二乘(OLS)回归所估计的结果，哪怕数据集很大。异常观察值也能对广义线性模型造成破坏性的损害，虽然这不是很常见。这进一步强化了发现并恰当地处理回归分析中的特异值/异常值(outlier)的重要意义。

将“现代”回归方法，如非参数回归，作为诊断工具整合进一般线性模型及广义线性模型的框架有很多好处(参见，如 Cook & Weisberg，1999；Fox，1997；Hastie、Tibshirani & Friedman 2001)。这些方法之所以被称为“现代方法”，是因为它们依赖密集计算，即在拟合大量回归的基础上计算出最终估计结果，它们能够揭示出只使用OLS估计时常常难以发现的大量问题——尤其是非线性问题，当然也包括其他残差方面的问题。只是在个人电脑运算速度已经极大提高的最近这段时间，社会统计学家才意识到这些方法的好处。