格致方法·定量研究系列　吴晓刚　主编

非参数回归：平滑散点图

[加] 约翰·福克斯（John Fox）　著
王骁　译　洪岩璧　校

- 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

43

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

非参数回归:平滑散点图

[加] 约翰·福克斯(John Fox) 著
王　骁 译　　洪岩璧 校

SAGE Publications ,Inc.

格致出版社　上海人民出版社

图书在版编目(CIP)数据

非参数回归:平滑散点图/(加)福克斯(Fox, J.)著;王骁译.—上海:格致出版社:上海人民出版社,2015

(格致方法·定量研究系列/吴晓刚主编)

ISBN 978-7-5432-2489-6

Ⅰ.①非… Ⅱ.①福… ②王… Ⅲ.①回归分析 Ⅳ.①O212.1

中国版本图书馆 CIP 数据核字(2015)第 038188 号

责任编辑 高 璇

美术编辑 路 静

格致方法·定量研究系列

非参数回归:平滑散点图

[加]约翰·福克斯 著

王骁 译 洪岩璧 校

出 版 世纪出版股份有限公司 格致出版社

世纪出版集团 上海人民出版社

(200001 上海福建中路 193 号 www.ewen.co)

格致出版

编辑部热线 021-63914988

市场部热线 021-63914081

www.hibooks.cn

发 行 上海世纪出版股份有限公司发行中心

印 刷 浙江临安曙光印务有限公司

开 本 920×1168 1/32

印 张 4.25

字 数 83,000

版 次 2015 年 4 月第 1 版

印 次 2015 年 4 月第 1 次印刷

ISBN 978-7-5432-2489-6/C·128

定价:22.00 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，翻译成中文，起初集结成八册，于2011年出版。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热烈欢迎。为了给广大读者提供更多的方便和选择，该丛书经过修订和校正，于2012年以单行本的形式再次出版发行，共37本。我们衷心感谢广大读者的支持和建议。

随着与SAGE出版社合作的进一步深化，我们又从丛书中精选了三十多个品种，译成中文，以飨读者。丛书新增品种涵盖了更多的定量研究方法。我们希望本丛书单行本的继续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

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”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,这两本书拖了多年还没有完成。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,与我的教学理念是相通的。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香港、台湾地区的二十几位

研究生参与了这项工程，他们当时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；哈佛大学社会学系博士研究生郭茂灿和周韵。

参与这项工作的许多译者目前都已经毕业，大多成为中国内地以及香港、台湾等地区高校和研究机构定量社会科学方法教学和研究的骨干。不少译者反映，翻译工作本身也是他们学习相关定量方法的有效途径。鉴于此，当格致出版社和 SAGE 出版社决定在“格致方法·定量研究系列”丛书中推出另外一批新品种时，香港科技大学社会科学部的研究生仍然是主要力量。特别值得一提的是，香港科技大学应用社会经济研究中心与上海大学社会学院自 2012 年夏季开始，在上海（夏季）和广州南沙（冬季）联合举办《应用社会科学研究方法研修班》，至今已经成功举办三届。研修课程设计体现“化整为零、循序渐进、中文教学、学以致用”的方针，吸引了一大批有志于从事定量社会科学研究的博士生和青年学者。他们中的不少人也参与了翻译和校对的工作。他们在

繁忙的学习和研究之余,历经近两年的时间,完成了三十多本新书的翻译任务,使得“格致方法·定量研究系列”丛书更加丰富和完善。他们是:东南大学社会学系副教授洪岩璧,香港科技大学社会科学部博士研究生贺光烨、李忠路、王佳、王彦蓉、许多多,硕士研究生范新光、缪佳、武玲蔚、臧晓露、曾东林,原硕士研究生李兰,密歇根大学社会学系博士研究生王骁,纽约大学社会学系博士研究生温芳琪,牛津大学社会学系研究生周穆之,上海大学社会学院博士研究生陈伟等。

陈伟、范新光、贺光烨、洪岩璧、李忠路、缪佳、王佳、武玲蔚、许多多、曾东林、周穆之,以及香港科技大学社会科学部硕士研究生陈佳莹,上海大学社会学院硕士研究生梁海祥还协助主编做了大量的审校工作。格致出版社编辑高璇不遗余力地推动本丛书的继续出版,并且在这个过程中表现出极大的耐心和高度的专业精神。对他们付出的劳动,我在此致以诚挚的谢意。当然,每本书因本身内容和译者的行文风格有所差异,校对未免挂一漏万,术语的标准译法方面还有很大的改进空间。我们欢迎广大读者提出建设性的批评和建议,以便再版时修订。

我们希望本丛书的持续出版,能为进一步提升国内社会科学定量教学和研究水平作出一点贡献。

吴晓刚

于香港九龙清水湾

序

在分析两个定量变量的关系时，我们要做的第一件事就是看一看散点图。这一视觉评估能够帮助我们判断函数形式。假设政治学家格温·格林(Gwen Greene)教授正用一个94国的样本研究国家的人口规模(x)和民选官员数(y)之间的关系。她的研究问题涉及变量相关关系是线性的还是其他什么形式？遗憾的是，散点图并不明确。她所看到的点云仅仅是一团缺乏清晰几何形状的点构成的云。对此，一个通常的回应是假定线性并使用最小二乘法(OLS)估计变量间的关系，但她知道这可能是个严重的错误。真正的关系可能是曲线型的，这意味着OLS结果会由于错误的模型设定而产生偏误。我们如何找到这条曲线？一种方法是从理论或前人研究中推导出来并对它建模，例如，或许我们需要估计一个二次方程。然而，设想一下在手头进行的研究中理论和前人研究可能会给出相互矛盾的建议。另一种方法就是通过系统性地探索数据来找到这条曲线(如果它的确存在)。后一种策略即引向福克斯教授在本书中所详细介绍的非参数简单回归技术。

非参数回归并不预设关于 x 和 y 的特定函数形式。恰恰相反,通过使用样本数据,它根据分组后的 x 值来计算不同 y 值的平均数。这些 y 值的平均数类似于被曲线连接的点被平滑为一条曲线。这条曲线可能呈现海浪状、蠕虫状或是其他不规则的形状,它表现出用一种更加精细的方式来描绘两个变量间的函数关系。如果散点图是通过这种方式被"平滑"的,那么通常所用的方法即局部加权回归的某个版本,后者通常被简称为"loess"。

由于对函数形式的搜索具有归纳性,很多曲线都是可能的。曲线的形状取决于"箱"(bin)定义、计算平均值的方法或是局部多项式回归的阶次。其他技术层面的问题则涉及核(kernel)估计、异常值的处理以及过度平滑(oversmoothing)。特别需要指出的是,如何确保曲线既不"太光滑"也不"太粗糙",不但是一门"科学"更是一门"艺术"。一旦曲线被确定下来,我们就可以沿曲线周围构造置信包迹(confidence envelope)并进一步进行假设检验。

正如其名称所暗示的,非参数回归的一个缺点是它无法得到对回归参数的估计。然而,该方法对适当函数形式的识别有助于产生无偏参数估计的一般性的理论模型设定。举例而言,比如格林教授在国家人口(x)和民选官员(y)之间找到了一条 loess 平滑曲线,其中随着 x 的增加,y 增加的速度越来越慢。这一曲线暗示变量间的关系是对数的,且可以通过在 OLS 中的变换将 y 表达成 $\log x$ 的函数。因此,平滑操作有助于发现能被进一步检验的更深层次的函数形式。福克斯教授书中的最后一章讨论了非参数回归对建立理论的作用,他将一般性的非线性问题与非参数回

归联系了起来。总而言之，这本心血之作继承了归纳科学令人尊敬的传统。它提醒我们，对数据深思熟虑的探索能够使我们获益匪浅。

迈克尔·S.刘易斯-贝克

目 录

第1章

什么是非参数回归？

回归分析通常是指用一个或几个预测变量(xs)的函数形式来描绘因变量(y)平均值的统计方法。假设有两个预测变量 x_1 和 x_2。这里把 y 基于预测变量的条件均值(也就是说,通过把预测变量限定在特定取值 x_1 和 x_2)记做 $\mu \mid x_1, x_2$,那么回归的主要目标就是通过样本来估计总体回归函数 $\mu \mid x_1, x_2 = f(x_1, x_2)$。同时,我们也可以在给定 xs 的情形下关注变量 y 条件分布的其他维度,例如 y 的中位数或者标准差。

正如通常所做的,回归分析假设 y 和 xs 之间存在线性关系,于是有

$$\mu \mid x_1, x_2 = f(x_1, x_2) = \alpha + \beta_1 x_1 + \beta_2 x_2$$

或者,等价地,

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

方程中误差项的均值为 0。我们通常也假设(至少是隐含地假设)除去均值之外,y 的条件分布在任何地方都相同,并且服从正态分布:

$$y \sim N(\alpha + \beta_1 x_1 + \beta_2 x_2, \sigma^2)$$

或者等价地,误差项服从具有相同方差的正态分布,$\varepsilon \sim$

$N(0, \sigma^2)$。最后,我们通常也假设观测值是被独立抽取的,因而y_i和$y_{i'}$(或者等价地,ε_i 和$\varepsilon_{i'}$)在 $i \neq i'$ 的条件下相互独立。

在这一整套假设满足的条件下就得到我们常用的线性最小二乘法回归。

上面列出的都是强假设,而且很多情况下会出错。例如,正如时间序列数据的典型案例,误差可能不是独立的,y 的条件方差(误差方差)也可能不相等,同时 y 的条件分布也可能不是正态分布而是重尾(heavy-tailed)或者倾斜的(skewed)分布。

非参数回归分析放松了线性假设,将其替换为平滑总体回归函数 $f(x_1, x_2)$所需的较弱的假设。放松线性假设的代价是需要更多的计算,以及在某些情形下的结果更难以理解。获益则是对回归函数更精确的估计。事实上,在一些实例中,盲目地使用线性假设会带来毫无意义的结果。

有些人可能会认为非参数回归显得“没理论”,从而反对在检验数据之前不给回归函数 $f(x_1, x_2)$指定函数形式的做法。我认为这种反对稍欠考虑:社会理论可能会告诉我们 y 取决于x_1 和 x_2,但却很少告诉我们它们的关系是线性的。而有效的统计数据分析的必要条件正是要能精确地描述数据。

本书的主题是非参数简单回归。这种方法中仅有单一应变量 y 和单一预测变量x,即 $y = f(x) + \varepsilon$。本书的姊妹篇讨论了广义非参数模型——例如,应变量是二分(有两个类别的)变量,以及非参数多元回归——含有多个预测变量。

一开始,非参数简单回归看起来可能会没什么用,因为

大多数关于回归分析的有趣应用会涉及多个预测变量。然而,有两个原因能说明非参数简单回归的用途:

1. 非参数简单回归通常被称做散点图平滑,因为在典型的应用中,这一方法会在 y 关于 x 的散点图上绘出一条经过若干点的平滑曲线。散点图是(或者说应当是)统计数据分析和演示中随处可见的要素,它既被用来初步查看回归数据,也被用来考察回归分析诊断图(参见第 7 章)。

2. 非参数简单回归的延伸构成了非参数多元回归的基础,同时也提供了被称做可加回归(additive regression)这一特定类型的非参数多元回归的组成元素。