格致方法·定量研究系列　吴晓刚　主编

样条回归模型

[美] 劳伦斯·C.马希（Lawrence C. Marsh）
戴维·R.科米尔（David R. Cormier）著
缪佳 译　许多多 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

63

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

样条回归模型

[美] 劳伦斯·C.马希(LawrenceC.Marsh)
戴维·R.科米尔(DavidR.Cormier) 著
缪佳 译　许多多 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

样条回归模型/(美)劳伦斯·C.马希(Lawrence C. Marsh),(美)戴维·R.科米尔(David R. Cormier)著;缪佳译.—上海:格致出版社:上海人民出版社,2017.8
(格致方法·定量研究系列)
ISBN 978-7-5432-2770-5

Ⅰ.①样… Ⅱ.①劳… ②戴… ③缪… Ⅲ.①样条函数-自回归模型 Ⅳ.①O241.5

中国版本图书馆CIP数据核字(2017)第155080号

责任编辑　张苗凤

格致方法·定量研究系列

样条回归模型

[美]劳伦斯·C.马希
戴维·R.科米尔　著
缪佳 译　许多多 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)
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　上海商务联西印刷有限公司
开　本　920×1168　1/32
印　张　4.25
字　数　68,000
版　次　2017年8月第1版
印　次　2017年8月第1次印刷

ISBN 978-7-5432-2770-5/C·183　　定价:30.00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

回归分析有很多类型，一种非常有用但常常被忽视的类型就是样条回归。样条回归和一系列概念有关，包括虚拟变量、时间计数、干预分析、中断时间序列、逐步线性回归等[在本丛书中，关于虚拟变量的讨论，请见哈迪(Hardy)的介绍(丛书编号 93*)；干预分析和中断的时间序列，见麦克道尔(McDowall)的介绍(丛书编号 21)]。我们假设有一个连续变量 Y，它随时间的发展轨迹因为某个事件或政策而发生了变化。举个具体的例子，Y 是一个国家每年领取福利的人数，在很多年里 Y 一直上升，之后由于实施了新的福利政策，Y 下降了。如果这种下降是突然的，就可以使用中断的时间序列模型，因为它可以反映出截距的变化。然而，如果领取福利的人数是缓慢降低而不是突然减少的，那么样条回归就更合适，因为它可以反映

* 该丛书编号为 SAGE 英文版系列编号。下同。——编者注

两条回归线在连接点处平滑的斜率变化，而避免了回归线中间出现断裂。

对一个简单模型，$Y=a+bT+cD(T-T_1)+e$，其中 Y 是每年领取福利的人数，T 是时间，以年计算，取值为 1，2，3，…，N。D 是虚拟变量，在福利政策改革前取值为 0，之后为 1。T_1 是改革以来的时间，也以年计算。对于福利政策改革之后的第 1，2，3，…年，$D(T-T_1)$ 分别等于 1，2，3，…，依次类推。假设“样条节点”，即改革发生的那一年是 1990 年，那么这个估计模型就会生成 1990 年之前和之后两条线性回归线，并且回归线之间没有突然跳跃。

当我们需要拟合回归线的弯曲或者变化时，样条回归是一种常用的方法。如果样条节点很少，并且我们事先已经知道它们的位置，这时的估计是最简单的。对于这种情况，马希(Marsh)和科米尔(Cormier)教授给出的例子是：在竞选的三个不同阶段，投票者的政党认同的变化。他们进而讨论更复杂的例子，即通过时间序列数据，分析 1890 年以来共和党和民主党的 11 次政权交替如何影响了债券利率的变化。同时他们还讨论了多项式模型在处理这类问题上的不足，后者面临着多重共线性的困境。

当样条节点位置未知时样条回归模型就更复杂了，这时需要用到非线性最小二乘估计。为了介绍估计程序，两位作者研究了三个(未知的)年龄点对个人的宗教虔诚度的影响。这个例子也说明：除了时间序列数据之外，样条

格致方法·定量研究系列

1. 社会统计的数学基础
2. 理解回归假设
3. 虚拟变量回归
4. 多元回归中的交互作用
5. 回归诊断简介
6. 现代稳健回归方法
7. 固定效应回归模型
8. 用面板数据做因果分析
9. 多层次模型
10. 分位数回归模型
11. 空间回归模型
12. 删截、选择性样本及截断数据的回归模型
13. 应用 logistic 回归分析（第二版）
14. logit 与 probit：次序模型和多类别模型
15. 定序因变量的 logistic 回归模型
16. 对数线性模型
17. 流动表分析
18. 关联模型
19. 中介作用分析
20. 因子分析：统计方法与应用问题
21. 非递归因果模型
22. 评估不平等
23. 分析复杂调查数据（第二版）
24. 分析重复调查数据
25. 世代分析（第二版）
26. 纵贯研究（第二版）
27. 多元时间序列模型
28. 潜变量增长曲线模型
29. 缺失数据
30. 社会网络分析（第二版）
31. 广义线性模型导论
32. 基于行动者的模型
33. 基于布尔代数的比较法导论
34. 微分方程：一种建模方法
35. 模糊集合理论在社会科学中的应用
36. 图解代数：用系统方法进行数学建模
37. 项目功能差异（第二版）
38. Logistic 回归入门
39. 解释概率模型：Logit、Probit 以及其他广义线性模型
40. 抽样调查方法简介
41. 计算机辅助访问
42. 协方差结构模型：LISREL 导论
43. 非参数回归：平滑散点图
44. 广义线性模型：一种统一的方法
45. Logistic 回归中的交互效应
46. 应用回归导论
47. 档案数据处理：生活经历研究
48. 创新扩散模型
49. 数据分析概论
50. 最大似然估计法：逻辑与实践
51. 指数随机图模型导论
52. 对数线性模型的关联图和多重图
53. 非递归模型：内生性、互反关系与反馈环路
54. 潜类别尺度分析
55. 合并时间序列分析
56. 自助法：一种统计推断的非参数估计法
57. 评分加总量表构建导论
58. 分析制图与地理数据库
59. 应用人口学概论：数据来源与估计技术
60. 多元广义线性模型
61. 时间序列分析：回归技术（第二版）
62. 事件史和生存分析（第二版）
63. 样条回归模型
64. 定序题项回答理论：莫坎量表分析

目 录

第1章

概述

样条回归听起来复杂，但实际上它只是附加了一些简单限制条件的虚拟变量模型。具体来讲，样条回归是有一个或多个连续性限制的虚拟变量模型。

例如，政治家的支持率可能在选举（改选）时最高，过后就会下降，这个趋势表现为一条斜率向下的回归线。到了某一时点，政治家会意识到应该努力拉升公众的支持率，为下一轮选举做准备。如果使用没有限制的虚拟变量，选举前和选举后的模型会有不同的截距和斜率，如图 1.1 所示。

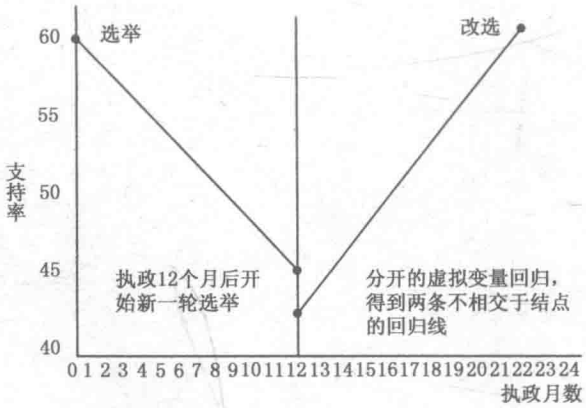

图 1.1　支持率的非限制虚拟变量回归

样条回归可以避免两条回归线之间出现突然“跳跃”(断裂)。在样条回归中,支持率的转折点由样条节点表示,这个节点将向下和向上的回归线连接起来,如图 1.2 所示。

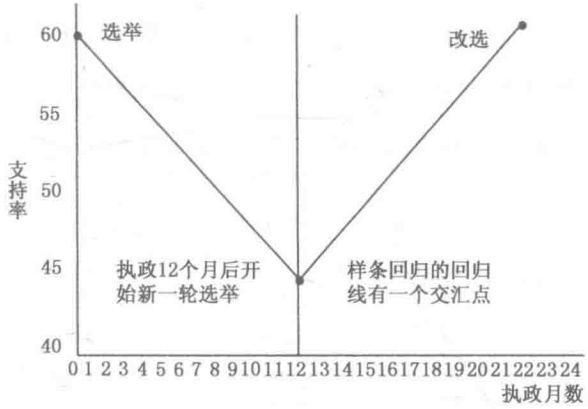

图 1.2 支持率的样条回归模型

在平代克和鲁宾菲尔德(Pindyck & Rubinfeld, 1998)编写的广受好评的教科书中,两位作者将此类型的样条回归模型称为逐段线性回归模型(piecewise linear regression model),并进行了简要介绍。休茨、梅森和陈(Suits, Mason & Chan, 1978)更直接且深入地讨论了样条回归。应用案例可以参见斯特劳津斯基(Strawczynski, 1998)应用逐段线性回归模型对税级距进行的研究。

在样条回归中,因变量的回归线(如支持率)的斜率会突然发生变化,但是回归线本身并不出现断裂或者“跳