分位回归与复杂分层结构数据分析

Quantile Regression & Complex Hierarchical Data Analysis

田茂再◎著

知识产权出版社
全国百佳图书出版单位

分位回归
与复杂分层结构数据分析

田茂再◉著

知识产权出版社
全国百佳图书出版单位

图书在版编目（CIP）数据

分位回归与复杂分层结构数据分析/田茂再著.
—北京：知识产权出版社，2015.6
ISBN 978-7-5130-2717-5

Ⅰ.①分… Ⅱ.①田… Ⅲ.①统计分析 Ⅳ.① O212.1

中国版本图书馆 CIP 数据核字(2014)第 089541 号

内容提要

具有复杂分层结构的数据在现实生活中很普遍，剖析这类数据，发现该类数据表象下的潜在规律对于统计学等科研领域很有意义。本书致力于介绍复杂分层数据分析的前沿知识，侧重于算法、仿真与实证研究，主要包括两大块内容：分位回归与分层—分位回归。

本书可作为统计学及其相关领域大学生、研究生的教学参考书，亦可供教师和科技人员参考。

责任编辑：江宜玲　　**责任出版**：刘译文

封面设计：回归线视觉传达

分位回归与复杂分层结构数据分析

田茂再◉著

出版发行：知识产权出版社 有限责任公司	网　　址：http://www.ipph.cn
社　　址：北京市海淀区马甸南村 1 号	邮　　编：100088
责编电话：010-82000860-8339	责编邮箱：jiangyiling@cnipr.com
发行电话：010-82000860 转 8101/8102	发行传真：010-82000893/82005070/82000270
印　　刷：三河市国英印务有限公司	经　　销：各大网上书店、新华书店及相关专业书店
开　　本：170mm×240mm 1/16	印　　张：23.5
版　　次：2015 年 6 月第 1 版	印　　次：2015 年 6 月第 1 次印刷
字　　数：468 千字	定　　价：128.00 元
ISBN 978-7-5130-2717-5	

前　言

很多分层数据具有以下分层结构：我们用变量来描述个体，而个体嵌套在更大单元里，形成金字塔形状。以教育方面的数据为例，学生被分成班级，班级嵌套在学校里。学校上面有社区，社区上面还有省、国家等。

自 20 世纪 70 年代以来，人们开始研究分层结构数据的统计模型。比如，作为对线性模型贝叶斯估计学术方面的贡献，Lindley & Smith (1972) 和 Smith (1973) 引入了分层线性模型 (Hierarchical Linear Model) 这一术语。然而，近年来分层模型在不同的领域有不同的称谓：在社会学研究里，叫作多水平模型 (Multilevel Model)，参见 Mason，Wong & Entwistle (1983)，Goldstein (1995)；生物统计上则称为混合效应模型 (Mixed-effects Model) 或者随机效应模型 (Random-effects Model)，参见 Elston (1962)，Laird (1982)，Longford (1987) 以及 Singer (1998)；计量经济学上称为随机系数回归模型 (Random-coefficient Regression Model)，参见 Rosenberg (1973) 和 Longford (1993)；在贝叶斯统计里，我们称之为条件独立分层模型 (Conditionally Independent Hierarchical Model)，参见 Kass & Steffey (1989)。一般的统计文献则称之为协方差成分模型 (Covariance Components Model)，参见 Dempster，Rubin & Tsutakawa (1981)。Hobert (2000) 给出了目前有关拟合分层模型计算方面的热点问题综述。

在上述所提到的各种模型背后，现有的分层模型理论主要关注的是在给定预测变量 X 的条件下，拟合响应变量 Y 的条件期望。尽管在很多应用中，这些理论能够应付了，然而它们却不能完全刻画响应变量在各分位点上的情况。例如，学校平均成绩有时候可能会隐藏一些涉及差生与优等生方面的问题，因为平均数本身不能对学生成绩提供一个“谱视”(Spectral View)。

分位回归 (Quantile Regression，QR) 方法，亦称分位数回归，产生于 30 年前。由于它能够全面刻画一个条件随机变量的各分位点随协变量的变化情况，所以近年来它逐渐发展成为一种综合的分析线性和非线性模型的统计方法。目前，有大量的文献是关于分位回归研究的。在本书中，我们充分利用了分层模拟与分位回归的优点，提出分层分位回归模型 (Hierarchical Quantile Regression Models)。这类模型具有如下特点：① 能够全面刻画出给定高维解释变量的条件下响应变量的各分位点情况；② 估计出来的系数向量，即边际效应，对于响应变量的离群观测值来说，是稳健的；③ 在不同分位点上潜在的不同解具有很有用的解释意义；④ 沿袭了分层模拟与分位回归模型二者所有的优点。

本书致力于介绍复杂分层数据分析前沿的知识，侧重于算法、仿真与实证研究，以给读者提供一些复杂分层数据的分位回归建模知识。

自 2004 年中国人民大学统计学院在全国首开《分位回归》课程以来，笔者一直担任本课程的主讲老师。本书的大部分材料在课堂上讨论过。本书在写作过程中，自始至终有以下硕士生、博士生参加过翻译、校正等工作：李远、周朋朋、范洁瑜、张宁、戴成、钱政超、石恒泽、周健、安姝静、陈博钰、范博文、范燕、姜春波、马维华、苏宇楠、张圆圆、陈彦靓、郭洁、康雁飞、荣耀华、王伟、罗幼喜、储昭雾、封达道、李兆媛、司世景、夏文涛、熊巍、何静、胡亚南、黄雅丽、李茜、刘甦倩、吕爽、朱倩倩、田玉柱、梁晓琳、马春桃、马绰欣、孟令宾、王榛、杨亚琦、张亚丽、李二倩、罗静、史普欣、王晓荷、袁梦、吴延科、晏振等。在此，我对他们表示衷心的感谢！

本书获得以下基金部分资助：国家自然科学基金 (No.11271368)，北京市社会科学基金重大项目 (No.15ZDA17)，教育部高等学校博士学科点专项科研基金 (No.20130004110007)，国家社会科学基金重点项目 (No.13AZD064)，中国人民大学科学研究基金 (中央高校基本科研业务费专项资金资助) 项目成果 (No.15XNL008)，教育部科学技术研究重点项目 (No.108120)，北京市社会科学基金项目 (No.12JGB051) 以及兰州商学院“飞天学者特聘计划”。同时感谢教育部人文社会科学重点研究基地中国人民大学应用统计研究中心的大力支持。

2014 年 5 月

于北京

目　　录

上篇　分位回归

第 1 章　分位回归引论 ····· 3

1.1　引言 ····· 3

1.1.1　分位数 ····· 3

1.1.2　分位回归 ····· 4

1.1.3　分位回归方法的演变 ····· 7

1.2　估计方法和算法 ····· 12

1.2.1　参数分位回归模型 ····· 12

1.2.2　Box-Cox 变换分位数模型 ····· 12

1.2.3　非参分位回归模型 ····· 13

1.2.4　窗宽选择 ····· 15

1.2.5　半参分位回归模型 ····· 16

1.2.6　两步法 ····· 17

1.3　分位回归应用领域 ····· 17

1.3.1　执行总裁年报酬与公司股本的市场价值关系 ····· 17

1.3.2　分位数恩格尔曲线 (Engel Curve) ····· 18

1.3.3　分位回归和婴儿体重的决定因素 ····· 20

1.3.4　医学中参考图表的应用 ····· 22

1.3.5　在生存分析方面的应用 ····· 23

1.3.6　风险值、分布尾部及分位数 ····· 24

1.3.7　经济 ····· 24

1.3.8　环境模型的应用 ····· 24

1.3.9　在检测异方差性上的应用 ····· 25

1.4　其他方面的进展 ····· 25

1.4.1　时间序列的分位回归 ····· 25

1.4.2　拟合优度 ····· 26

1.4.3　贝叶斯分位回归 ····· 27

1.5　软件和标准误差 ····· 27

1.6 文献介绍 ····· 28
第 2 章 线性分位回归模拟 ····· 30
2.1 基本概念 ····· 30
2.1.1 基于条件分位函数的定义 ····· 30
2.1.2 基于分位回归模型的定义 ····· 30
2.1.3 基于损失函数的定义 ····· 30
2.1.4 基于非对称拉普拉斯密度的定义 ····· 31
2.2 家庭背景因素的影响 ····· 31
2.3 数据 ····· 32
2.4 估计结果 ····· 34
2.4.1 10 年级的影响估计 ····· 34
2.4.2 11 年级的影响估计 ····· 35
2.4.3 12 年级的影响估计 ····· 36
2.5 置信区间和相关解释 ····· 39
2.5.1 哪一个是最好的? 双亲、单亲还是没有父母 ····· 39
2.5.2 为什么我们要关注兄弟姐妹关系 ····· 40
2.5.3 父亲和母亲之间的影响的区别是什么 ····· 40
2.5.4 性别上有差异吗 ····· 40
2.5.5 表现差距在哪里 ····· 40
2.5.6 语言问题是很严重的问题吗 ····· 41
2.5.7 本地学生从数学教学中获益了吗 ····· 41
2.6 结论 ····· 41
2.7 文献介绍 ····· 42
第 3 章 非参数分位回归模拟 ····· 43
3.1 稳健局部逼近 ····· 43
3.1.1 介绍 ····· 43
3.1.2 LAM 估计的相合性 ····· 44
3.1.3 LAM 估计的渐近分布 ····· 46
3.1.4 $I = 2$ 条件下关于 K 和 β 的最优估计 ····· 46
3.1.5 文献介绍 ····· 48
3.2 非参数函数估计 ····· 48
3.2.1 引言 ····· 48
3.2.2 渐近性质 ····· 50
3.2.3 百分位回归和预测区间 ····· 51
3.2.4 文献介绍 ····· 53

3.3 局部线性分位回归 ····· 53
3.3.1 引言 ····· 53
3.3.2 局部线性检验函数的最小化 ····· 56
3.3.3 局部线性双核平滑 ····· 60
3.3.4 实际性能 ····· 63
3.3.5 文献介绍 ····· 66
3.4 教育数据分析 ····· 67
3.4.1 数据 ····· 68
3.4.2 方法 ····· 69
3.4.3 科学成绩 ····· 70
3.4.4 数学成绩 ····· 73
3.4.5 科学成绩和数学成绩的关系 ····· 75
3.4.6 文献介绍 ····· 77
第 4 章 适应性分位回归模拟 ····· 78
4.1 局部常数适应性分位回归 ····· 78
4.1.1 引言 ····· 78
4.1.2 适应性估计 ····· 79
4.1.3 实现 ····· 81
4.1.4 理论性质 ····· 82
4.1.5 蒙特卡洛研究 ····· 83
4.1.6 不同方法的比较 ····· 87
4.1.7 局部适应性窗宽的自动选择 ····· 88
4.1.8 应用 ····· 91
4.1.9 文献介绍 ····· 91
4.2 局部线性适应性分位回归 ····· 92
4.2.1 介绍 ····· 92
4.2.2 局部线性适应性估计 ····· 93
4.2.3 算法 ····· 95
4.2.4 理论性质 ····· 96
4.2.5 蒙特卡洛模拟 ····· 97
4.2.6 文献介绍 ····· 99
第 5 章 可加性分位回归模拟 ····· 100
5.1 高维协变量下可加条件分位回归 ····· 100
5.1.1 引言 ····· 100
5.1.2 方法 ····· 102

5.1.3 渐近性质 ····· 105
5.1.4 与后拟合方法在数值表现上的比较 ····· 108
5.1.5 例子 ····· 111
5.1.6 文献介绍 ····· 115
5.2 可加分位回归的非参数估计 ····· 115
5.2.1 介绍 ····· 116
5.2.2 估计量的正式描述 ····· 118
5.2.3 一个经验例子 ····· 119
5.2.4 渐近结果 ····· 121
5.2.5 蒙特卡洛实验 ····· 125
5.2.6 文献介绍 ····· 127
第 6 章 变系数分位回归模拟 ····· 128
6.1 适应性变系数分位回归 ····· 128
6.1.1 引言 ····· 128
6.1.2 自适应估计 ····· 129
6.1.3 理论性质 ····· 134
6.1.4 实证例子 ····· 136
6.1.5 文献介绍 ····· 141
6.2 异方差变系数分位回归 ····· 141
6.2.1 引言 ····· 141
6.2.2 局部线性 CQR-AQR 估计 ····· 143
6.2.3 局部二次 CQR-AQR 估计 ····· 147
6.2.4 窗宽选择 ····· 148
6.2.5 假设检验 ····· 149
6.2.6 数值模拟 ····· 150
6.2.7 经验应用 ····· 157
6.2.8 局部 m 次多项式 CQR-AQR 估计 ····· 159
6.2.9 文献介绍 ····· 161
第 7 章 单指数分位回归模拟 ····· 162
7.1 引言 ····· 162
7.2 模型与估计 ····· 163
7.2.1 模型与局部线性估计 ····· 163
7.2.2 带宽选择 ····· 166
7.3 大样本性质 ····· 167
7.3.1 非参部分的渐近性 ····· 167

7.3.2 参数部分的渐近性 ····· 168
7.4 数值研究 ····· 169
7.4.1 模拟 ····· 169
7.4.2 波士顿房价数据应用 ····· 173
7.5 文献介绍 ····· 176
第 8 章 分位自回归模拟 ····· 177
8.1 引言 ····· 177
8.2 模型 ····· 178
8.2.1 模型界定 ····· 178
8.2.2 分位自回归过程的性质 ····· 179
8.3 估计 ····· 181
8.4 分位单调性 ····· 183
8.5 分位自回归过程的统计推断 ····· 186
8.5.1 回归 Wald 检验过程与相关检验 ····· 187
8.5.2 非对称动态性检验 ····· 187
8.6 蒙特卡洛 ····· 189
8.7 实证运用 ····· 191
8.7.1 失业率 ····· 192
8.7.2 汽油零售价的动态性 ····· 192
8.8 文献介绍 ····· 194
第 9 章 复合分位回归模拟 ····· 195
9.1 复合分位回归与模型选择 ····· 195
9.1.1 介绍和动机 ····· 195
9.1.2 复合分位回归 ····· 197
9.1.3 渐近相对有效性 ····· 198
9.1.4 CQR-Oracular 估计量 ····· 203
9.1.5 模拟研究 ····· 204
9.1.6 文献介绍 ····· 205
9.2 局部复合分位回归 ····· 205
9.2.1 引言 ····· 205
9.2.2 回归函数的估计 ····· 206
9.2.3 导数的估计 ····· 210
9.2.4 数值比较和例子 ····· 214
9.2.5 局部 p 阶多项式复合分位回归光滑和证明 ····· 220
9.2.6 讨论 ····· 221

9.2.7 文献介绍 ····· 222
第 10 章 高维分位回归模拟 ····· 223
10.1 引言 ····· 223
10.2 非凸惩罚的分位回归 ····· 224
10.2.1 方法 ····· 224
10.2.2 差分凸规划及充分局部最优性条件 ····· 226
10.2.3 渐近性质 ····· 226
10.3 模拟与实际数据例子 ····· 229
10.3.1 模拟研究 ····· 230
10.3.2 应用 ····· 232
10.4 文献介绍 ····· 236
第 11 章 贝叶斯分位回归模拟 ····· 237
11.1 引言 ····· 237
11.2 非对称拉普拉斯分布 ····· 238
11.3 贝叶斯分位回归 ····· 239
11.4 参数的不合适先验 ····· 240
11.5 应用 ····· 240
11.5.1 模拟数据 ····· 240
11.5.2 免疫球蛋白 IgG ····· 242
11.5.3 烟囱损失 ····· 242
11.6 文献介绍 ····· 244

下篇 分层分位回归模拟

第 12 章 分层样条分位回归模拟 ····· 247
12.1 引言 ····· 247
12.2 条件分位函数的非参估计 ····· 248
12.3 回归分位数模型的 Wald 检验 ····· 250
12.4 条件分位分层模型及其在家庭用电量需求上的应用 ····· 252
12.4.1 第一阶段: 家庭需求周期的时间序列模型 ····· 252
12.4.2 第二阶段: 需求周期的横截面模型 ····· 253
12.4.3 条件分位数分层模型 ····· 254
12.5 数据的描述 ····· 255
12.5.1 第一阶段结果 ····· 256
12.5.2 第二阶段结果 ····· 257

12.6 文献介绍 ····· 262
第 13 章 分层线性分位回归模拟 ····· 264
13.1 引言 ····· 264
13.2 分层分位回归模型 ····· 264
13.3 EQ 算法 ····· 265
13.3.1 Q 步 ····· 265
13.3.2 E 步 ····· 266
13.3.3 迭代 ····· 267
13.3.4 初始值选取的基本方法 ····· 267
13.4 渐近性质 ····· 267
13.5 真实数据分析举例 ····· 269
13.5.1 数据描述 ····· 269
13.5.2 分位回归 ····· 269
13.5.3 两水平分层分位回归模型 ····· 270
13.5.4 部分结果 ····· 272
13.6 文献介绍 ····· 274
第 14 章 分层半参数分位回归模拟 ····· 275
14.1 介绍 ····· 275
14.2 模型和估计 ····· 276
14.2.1 研究 J 所学校 SES 成绩之间的关系 ····· 277
14.2.2 母亲讲话对孩子词汇量的影响 ····· 278
14.3 渐近结果 ····· 282
14.4 模拟分析 ····· 283
14.4.1 误差为多元柯西分布的层次线性模型 ····· 283
14.4.2 具有异方差的层次非参分位回归模型 ····· 284
14.5 实际数据例子 ····· 286
14.6 文献介绍 ····· 289
第 15 章 复合分层线性分位回归模拟 ····· 290
15.1 介绍 ····· 290
15.2 模型 ····· 291
15.3 估计 ····· 292
15.3.1 CQ 步 ····· 292
15.3.2 E 步 ····· 292
15.3.3 迭代 ····· 293
15.4 渐近性质 ····· 294

15.4.1　误差项为正态分布 ····· 294
15.4.2　误差项分布非正态 ····· 295
15.5　模拟 ····· 296
15.5.1　误差项为正态分布 ····· 296
15.5.2　误差项为柯西分布 ····· 296
15.5.3　离群点 ····· 297
15.5.4　选择最优 K ····· 298
15.6　实证部分 ····· 299
15.6.1　描述数据 ····· 299
15.6.2　多水平模型中的数据分析 ····· 299
15.6.3　结果 ····· 300
15.7　文献介绍 ····· 302
第 16 章　复合分层半参数分位回归模拟 ····· 303
16.1　介绍 ····· 303
16.2　模型 ····· 304
16.2.1　第一层单元内部模型 ····· 304
16.2.2　第二层单元之间模型 ····· 304
16.3　估计与算法 ····· 305
16.4　渐近性质 ····· 306
16.5　模拟研究 ····· 308
16.5.1　对于不同的误差项分布 ····· 308
16.5.2　对于 Y 存在异常值的情况 ····· 310
16.5.3　函数及其导数估计 ····· 311
16.6　实际数据分析 ····· 312
16.6.1　第一次层模型 ····· 314
16.6.2　第二次层模型 ····· 314
16.7　文献介绍 ····· 315
参考文献 ····· 317

上篇

分位回归

第 1 章　分位回归引论

1.1　引　　言

分位回归由 Koenker & Bassett (1978) 提出，它可以看作是将经典的最小二乘方法从估计条件均值模型扩展到估计条件分位函数组合的模型。一个重要的特殊情况就是中位数回归估计量，它是最小化绝对误差的和。其他的条件分位函数的估计方法是通过计算最小化绝对误差的非对称加权和。

1.1.1　分位数

1.1.1.1　总体无条件分位数

令随机变量 Y 的累积分布函数为 $F(y)$，则它的 τ 阶分位数 (无条件的) 定义为

$$Q_\tau(Y) = \arg\inf\left[y \in \mathbb{R};\ F(y) \geqslant \tau\right] \quad (0 < \tau < 1)$$

若将分布函数 $F(x)$ 的逆定义为

$$F_Y^{-1}(\tau) = \inf\left[y \in \mathbb{R};\ F(y) \geqslant \tau\right]$$

则

$$Q_\tau(Y) = F_Y^{-1}(\tau)$$

其实，分位数这个术语与百分数是同义的；中位数是分位数一个最熟知的例子。通常，用样本中位数作为总体中位数 m 的一个估计量。总体中位数是一个量，它将分布分割成两部分。如果对于总体分布来说，一个随机变量 Y 是可以被测量的，则 $P(Y \leqslant m) = P(Y \geqslant m) = 1/2$。特别地，对于一个连续型随机变量，$m$ 是等式 $F(m) = 1/2$ 的一个解。其中，$F(y) = P(Y \leqslant y)$ 为累积分布函数。由于只有少数人赚取巨额的工资，所以工资的分布是典型右偏的。因此对于典型的工资，与均值相比，样本中位数是一个更好的概括。

更一般的，25%样本分位数可以被定义为将数据分割成 1/4 和 3/4 两部分的值。反过来，可以定义为 75%样本分位数。相应的，连续情形中总体的下 1/4 分位数和上 3/4 分位数各自为等式 $F(y) = 1/4$ 和 $F(y) = 3/4$ 的解。一般来说，对于一个比例 $\tau(0 < \tau < 1)$，在连续情形中，F 的 100τ%分位数 (等价的，第 100τ 的百分位数) 是 $F(y) = \tau$ 的解 y，我们假定解是唯一的。

1.1.1.2　样本无条件分位数

令 $Y_{(1)} \leqslant Y_{(2)} \leqslant \cdots \leqslant Y_{(n)}$ 表示一组来自总体 $F(y)$ 的随机样本 $\{Y_i\}_{i=1}^n$ 的顺序统计量。$F(y)$ 的传统估计方法是非参数密度估计所得的经验分布函数 $F_n(y)$，则 τ 阶分位数 $F^{-1}(\tau)(0<\tau<1)$ 的经验估计为

$$F_n^{-1}(\tau) = X_{([n\tau])}$$

式中：符号 $[\cdot]$ 为 $\cdot$ 的取整。

我们知道样本中位数可以被定义为一个排了序的数据集合的中间值 (或是两个中间值的一半)，也就是说，样本中位数将数据分成两部分，每部分的数据个数是相等的。

在一次标准考试中，如果一个学生的成绩处在 τ 分位数，那就是说该生表现得要比 τ (例如 80%) 比例的学生好，同时比 $(1-\tau)$ (例如 20%) 的学生差。所以，一半的学生表现得比中位数上的学生好，而另一半则表现得比中位数差。类似地，四分位数将总体分为 4 段，在每一段中所占比例是相同的。五分位数将总体分为 5 段；十分位数则将总体分为 10 段。在一般情况下，分位数又叫作百分位数，有时又称作分位数。分位回归由 Koenker & Bassett (1978) 提出，以寻求扩展这些思想去估计条件的分位函数模型。模型中响应变量条件分布的分位数标示为观察到的协变量的函数。

1.1.1.3　总体条件分位数

设有随机向量 (X, Y)，其中 Y 在给定 $X=x$ 的情况下的条件累积分布函数为 $F_{Y|X=x}(y|x)$，则将该条件随机变量 $Y|X=x$ 的 τ 阶分位数 (条件的) 定义为

$$Q_\tau(Y|X=x) = \arg\inf\left[y \in \mathbb{R};\ F(y|x) \geqslant \tau\right] \quad (0<\tau<1)$$

1.1.2　分位回归

我们知道，均值回归研究的是给定解释变量后响应变量的平均变化趋势，而分位回归则试图全面刻画条件随机变量的各分位点随解释变量的变化情况。图 1-1 粗略地描绘了人类在其历史长河中身体各部位高度的变化分位曲线图，可以看出踝关节、膝关节、髋关节、下颌以及整个身高的变化并非呈直线趋势。同时，我们也注意到中位数回归曲线与均值回归曲线接近。

从模型角度来讲，假定我们有样本序列 $\{(X_i, Y_i), (i=1, \cdots, n)\}$ 满足下列回归模型，即

$$Y = m(X) + \epsilon, \quad X \in \mathbb{R}^d$$

式中：X_i $(i=1, \cdots, n)$ 为固定设计点。