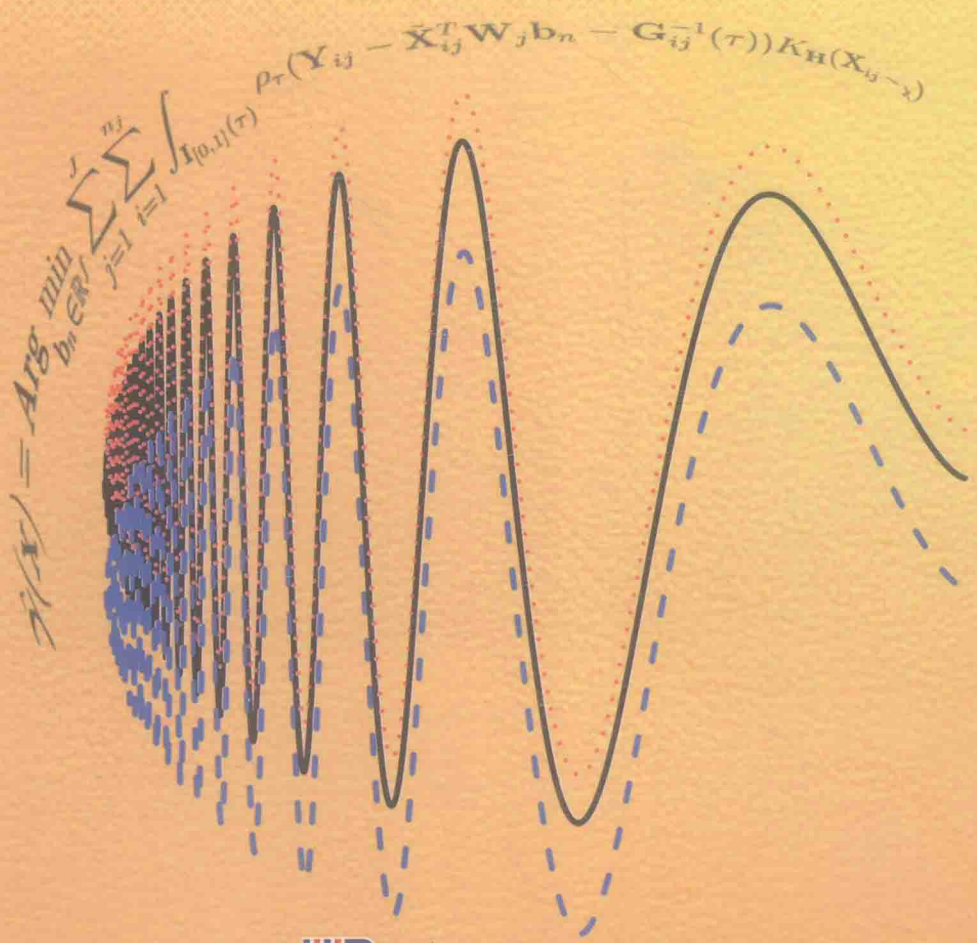
高等分层分位回归建模理论

田茂再 著



科学出版社

高等分层分位回归建模理论

田茂再　著

科学出版社
北京

内 容 简 介

具有复杂时空结构的数据很普遍，这类分层结构数据的分析方法很快应用到各研究领域. 本书全面、系统、严格地阐明分层分位回归建模理论与方法，并尽力反映复杂分层数据分析国际前沿研究. 内容涉及分层线性分位回归模型、分层广义线性分位回归模型、分层非线性分位回归模型、分层半参数分位回归模型等该领域前沿课题.

本书可作为统计学及其相关领域的本科生、研究生的教学参考书，也可供教师和科技人员参考.

图书在版编目(CIP)数据

高等分层分位回归建模理论/田茂再著. —北京：科学出版社，2014.6

ISBN 978-7-03-040875-4

Ⅰ. ①高… Ⅱ. ①田… Ⅲ. ①数理统计-统计模型 Ⅳ. ①O212

中国版本图书馆 CIP 数据核字(2014)第 114073 号

责任编辑：王丽平／责任校对：胡小洁

责任印制：肖 兴／封面设计：陈 敬

安徽大学图书馆 藏书 www.lib.ahu.edu.cn

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京通州皇家印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2015 年 1 月第 一 版 开本：720 × 1000 1/16

2015 年 1 月第一次印刷 印张：21 3/4

字数：430 000

定价：118.00 元

(如有印装质量问题，我社负责调换)

前　言

具有复杂分层结构特征的数据普遍存在于现实世界中. 离开数据的这一 “时空性”, 常常会使传统的统计分析方法效果不佳, 甚至失效. 分层结构数据分析是近二十年来迅速发展起来的一个统计学分支, 在各学科领域的应用越来越受到重视, 相关的研究显得异常活跃, 以至于文献中出现了各种各样的称谓. 在社会学研究方面, 这些模型指分层线性模型 (Goldstein,1995; Mason et al.,1983); 在生物测定学方面, 这些术语普遍指混合型效应模型和随机效应模型 (Elston and Grizzle, 1962; Laird and Ware, 1982; Singer, 1998); 在计量经济学文献中, 也称之为随机系数模型 (Rosenberg, 1973; Longford, 1993); 而在统计文献中则称之为协方差成分模型 (Dempster et al., 1981; Longford, 1987) 以及多水平模型等.

此外, Koenker 和 Bassett (1978) 先提出了分位回归模型的概念. 分位回归是一种统计方法, 它旨在对条件分位函数进行统计推断. 正如基于误差平方和最小化的经典线性回归方法能估计条件均值函数一样, 分位回归方法为我们提供了一种估计条件分位函数的机制. 有关分位回归的优点可以粗略地概括如下：① 给定一组预测变量之后, 它能全面刻画响应变量的整个条件分布; ② 分位回归模型有线性规划代理 (LP), 这使得估计简便; ③ 就像 LAD 这一特例一样, 分位回归的目标函数是加权的绝对偏差和, 所以它能给出一个稳健的位置测度, 因此被估计的系数向量对响应变量的离群点 (outliers) 不敏感, 特别是当误差项服从非正态的时候, 分位回归估计量要比最小二乘估计量更为有效. 最近分位回归取得了长足发展. 将分位回归方法应用于分层结构数据分析, 形成了分层分位回归建模的基本思想.

但是, 目前非常缺少全面、系统介绍分层分位回归建模理论与方法方面的书, 更谈不上能找到全力反映复杂分层数据分析国际前沿研究的专著了. 本书正朝着这个目标努力, 用严格的数学语言对分层分位模拟的现代面貌做较为详细的介绍.

本书的特点之一是内容全新、理论性强, 介绍近年来国际上关于分层分位回归建模理论与方法的许多新成果. 其中不乏作者在该领域研究中取得的一些成果. 另一个特点是取材全部来自于国际一流学术期刊杂志上的代表性文章, 信息量大、内容权威.

自 2004 年中国人民大学统计学院在全国首开分层模型课程以来, 本书作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, 自始至终有下面的硕士生、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、储昭霁、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. 对此, 对他们表示衷心的感谢!

本书得到下面基金的部分资助：国家自然科学基金 (No.11271368), 北京市哲学社会科学规划项目 (No.12JGB051), 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项目 (No.10XNL018,10XNK025), 全国统计科研计划项目 (No. 2011LZ031), 以及甘肃省“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.

由于作者水平有限, 疏漏和不足在所难免, 甚望读者批评指正!

田茂再

2013 年 6 月

此为试读,需要完整PDF请访问: www.ertongbook.com

目　录

上篇　分层结构数据均值建模理论、方法

第 1 章　分层线性模型 ····· 3

1.1　引言 ····· 3

1.1.1　背景 ····· 3

1.1.2　复杂数据概念 ····· 4

1.1.3　传统模型 ····· 5

1.1.4　主要参考文献 ····· 7

1.2　极大似然法 ····· 7

1.2.1　引言 ····· 7

1.2.2　EM 算法的定义 ····· 10

1.2.3　一般性质 ····· 14

1.2.4　主要参考文献 ····· 20

1.3　EM 算法 ····· 20

1.3.1　介绍 ····· 20

1.3.2　协方差已知情况下的理论 ····· 21

1.3.3　方差和协方差估计 ····· 24

1.3.4　计算 ····· 26

1.3.5　主要参考文献 ····· 33

1.4　迭代广义最小二乘法 ····· 33

1.4.1　引言 ····· 33

1.4.2　基本模型 ····· 34

1.4.3　估计 ····· 36

1.4.4　误差方差、协方差及误差协方差矩阵逆 ····· 38

1.4.5　随机系数 ····· 40

1.4.6　参数限制 ····· 41

1.4.7　解释变量有测量误差的估计量 ····· 41

1.4.8　讨论 ····· 43

1.4.9　主要参考文献 ····· 44

1.5　得分算法 ····· 44

1.5.1 引言 ····· 44
1.5.2 模型 ····· 45
1.5.3 对数似然函数 ····· 47
1.5.4 二水平嵌套 ····· 49
1.5.5 期望信息阵、雅可比行列式以及对数似然 ····· 51
1.5.6 EM 算法 ····· 52
1.5.7 多于两水平嵌套 ····· 53
1.5.8 主要参考文献 ····· 55
1.6 Newton-Raphson 算法 ····· 56
1.6.1 引言 ····· 56
1.6.2 计算方法 ····· 58
1.6.3 对数似然的导数 ····· 59
1.6.4 矩阵分解 ····· 62
1.6.5 参数 σ 与 D 的估计 ····· 65
1.6.6 讨论与延伸 ····· 67
1.6.7 主要参考文献 ····· 69
1.7 贝叶斯法 ····· 69
1.7.1 引言 ····· 70
1.7.2 三种情况 ····· 77
1.7.3 协方差结构未知时的估计 ····· 81
1.7.4 协方差结构未知的例子 ····· 82
1.7.5 多元回归方程间的可交换性 ····· 83
1.7.6 多元回归方程中的可交换性 ····· 86
1.7.7 主要参考文献 ····· 87
第 2 章 分层广义线性模型 ····· 88
2.1 模型 ····· 88
2.1.1 介绍 ····· 88
2.1.2 分层广义线性模型 ····· 89
2.1.3 典则连接模型 ····· 94
2.1.4 对数连接模型 ····· 94
2.1.5 典则连接模型 ····· 95
2.1.6 对数连接模型 ····· 96
2.1.7 最大 h 似然估计的性质 ····· 97
2.1.8 估计过程 ····· 103
2.1.9 推广 ····· 108

2.1.10 讨论 ····· 113
2.1.11 主要参考文献 ····· 114
2.2 抽样方法 ····· 115
2.2.1 引言 ····· 115
2.2.2 随机效应广义线性模型 ····· 117
2.2.3 贝叶斯公式 ····· 117
2.2.4 Gibbs 抽样 ····· 118
2.2.5 条件分布 ····· 119
2.2.6 讨论 ····· 122
2.2.7 主要参考文献 ····· 122
第 3 章 分层非线性模型 ····· 124
3.1 二阶广义估计方程 ····· 124
3.1.1 引言 ····· 124
3.1.2 模型 ····· 125
3.1.3 估计 ····· 126
3.1.4 条件方差–协方差的结构 ····· 129
3.1.5 惩罚尾似然和惩罚扩展最小二乘的关系 ····· 130
3.1.6 渐近性质 ····· 131
3.1.7 讨论 ····· 139
3.1.8 主要参考文献 ····· 140
3.2 混合估计 ····· 141
3.2.1 引言 ····· 141
3.2.2 3 个估计量 ····· 142
3.2.3 混合估计 ····· 145
3.2.4 渐近理论 ····· 148
3.2.5 推广 ····· 157
3.2.6 讨论 ····· 158
3.2.7 主要参考文献 ····· 159
第 4 章 分层半参数模型 ····· 160
4.1 分层半参数非线性模型 ····· 160
4.1.1 引言 ····· 160
4.1.2 半参非线性混合效应模型 ····· 162
4.1.3 估计 ····· 165
4.1.4 计算 ····· 168
4.1.5 统计推断 ····· 171

4.1.6 结论 ····· 172
4.1.7 主要参考文献 ····· 173
4.2 联合模型 ····· 173
4.2.1 背景 ····· 174
4.2.2 模型与估计方法 ····· 175
4.2.3 渐近性质 ····· 178
4.2.4 稳健性 ····· 188
4.2.5 讨论 ····· 188
4.2.6 主要参考文献 ····· 189

下篇　分层结构数据的分位回归模型理论、方法

第 5 章　分位回归引论 ····· 193
5.1 引言 ····· 193
5.1.1 分位数 ····· 193
5.1.2 分位回归 ····· 194
5.1.3 分位回归方法的演变 ····· 197
5.2 估计方法和算法 ····· 202
5.2.1 参数分位回归模型 ····· 202
5.2.2 Box-Cox 变换分位数模型 ····· 202
5.2.3 非参分位回归模型 ····· 203
5.2.4 窗宽选择 ····· 205
5.2.5 半参分位回归模型 ····· 206
5.2.6 两步法 ····· 207
5.3 应用领域 ····· 207
5.3.1 报酬与市场价值 ····· 207
5.3.2 分位数 Engel 曲线 ····· 209
5.3.3 婴儿体重的决定因素 ····· 210
5.3.4 医学中参考图表 ····· 212
5.3.5 生存分析 ····· 213
5.3.6 风险值与分布尾部 ····· 214
5.3.7 经济 ····· 214
5.3.8 环境 ····· 214
5.3.9 异方差性检测 ····· 215
5.4 其他方面的进展 ····· 215

5.4.1 时间序列 ····· 215
5.4.2 拟合优度 ····· 216
5.4.3 贝叶斯分位回归 ····· 217
5.5 主要参考文献 ····· 217
第 6 章 分层样条分位回归模型 ····· 219
6.1 引言 ····· 219
6.2 非参估计 ····· 220
6.3 Wald 型检验 ····· 222
6.4 实际应用 ····· 225
6.4.1 第一层: 时间序列模型 ····· 225
6.4.2 第二层: 横截面模型 ····· 226
6.4.3 条件分位数分层模型 ····· 227
6.5 结论 ····· 228
6.6 主要参考文献 ····· 229
第 7 章 分层线性分位回归模型 ····· 230
7.1 引言 ····· 230
7.2 模型界定 ····· 231
7.3 EQ 算法 ····· 232
7.3.1 Q 步 ····· 232
7.3.2 E 步 ····· 233
7.3.3 迭代 ····· 233
7.3.4 初始值选取 ····· 234
7.4 大样本性质 ····· 234
7.5 主要参考文献 ····· 240
第 8 章 分层半参数分位回归模型 ····· 241
8.1 分层半参数分位回归 ····· 241
8.2 引言 ····· 241
8.3 模型和估计 ····· 242
8.4 渐近结果 ····· 247
8.5 结论 ····· 254
8.6 主要参考文献 ····· 254
第 9 章 复合分层线性分位回归模型 ····· 255
9.1 复合分层线性分位回归 ····· 255
9.2 引言 ····· 255
9.3 模型 ····· 256

9.4 估计 ····· 257
9.5 大样本性质 ····· 259
9.5.1 误差项为正态分布情形 ····· 259
9.5.2 误差项分布非正态情形 ····· 263
9.6 讨论 ····· 264
9.7 主要参考文献 ····· 264
第 10 章 复合分层半参数分位回归模型 ····· 266
10.1 复合分层半参数分位回归 ····· 266
10.2 引言 ····· 266
10.3 模型 ····· 267
10.4 估计与算法 ····· 268
10.5 大样本性质 ····· 269
10.6 讨论 ····· 275
10.7 主要参考文献 ····· 275
第 11 章 分层空间模型的逆问题 ····· 277
11.1 分位耦合 ····· 277
11.2 引言 ····· 277
11.3 分位耦合 ····· 278
11.4 分层序列空间模型 ····· 279
11.4.1 模型 ····· 279
11.4.2 耦合步骤 ····· 280
11.4.3 自适应齐性诊断 ····· 283
11.5 非渐近神谕不等式 ····· 284
11.5.1 估计 ····· 284
11.5.2 权重序列的神谕性质 ····· 285
11.5.3 数据驱动的权重序列的选择 ····· 286
11.6 主要参考文献 ····· 291

参考文献 ····· 293
索引 ····· 331

此为试读，需要完整PDF请访问：www.ertongbook.com

上　篇

分层结构数据均值建模理论、方法

第 1 章　分层线性模型

1.1 引　　言

1.1.1 背景

1. *分层结构数据很普遍*

数据存在于特定的时间和空间中, 其表现形式通常是复杂的分层结构, 这是一种非常普遍的现象. 比如, 公司在制定决策以便提高劳动生产率方面, 显然工人和公司都是分析的对象. 对这两个层次的变化都必须进行考量. 其实这样的数据就有着一种分层结构: 工人嵌套在公司里. 又如, 在研究国家经济的发展与影响生育率的教育时, 家庭和国家都是研究的对象, 前者嵌套在后者之中, 这基本的数据结构也是分层的. 再给出一个例子: 关于教育方面的数据, 学生被分成班级, 班级嵌套在学校里, 学校上面有社区, 社区上面还有市、省、国家等.

具有分层结构的数据是一种普遍现象. 随着科学技术的飞速发展, 当今世界许多科学研究领域主要面临的挑战是急剧增长的高维多元复杂分层数据, 这种类型的数据普遍存在. 有关这种类型的数据的有效分析无论是在理论研究方面还是在经验研究方面都引起了广泛的关注. 本书着力研究如下 6 大类型的高维空间里的分层结构数据: ① 空间分层数据 (hierarchical data); ② 时间纵向数据 (longitudinal data); ③ 重复测量数据 (repeated measurement data); ④ 广义聚类数据 (generalized clustered data); ⑤ 名义分类数据 (nominal categorical data); 以及 ⑥ 有序分类数据 (ordinal categorical data) 等. 如何从中挖掘出有用的信息, 找出数据掩盖下的事物存在与发展的基本规律, 促进统计学学科的发展, 推动若干重要的相关领域及某些科学前沿取得突破, 这些正是本书研究的目的.

2. *模型名称*

由于分层数据分析在各学科领域越来越受到重视, 所以相关的研究显得异常活跃, 文献中出现了各种各样的称谓. 在社会学研究方面, 这些模型指分层线性模型 (Goldstein, 1995; Mason et al., 1983); 在生物测定学方面, 这些术语普遍指混合型效应模型和随机效应模型 (Elston 和 Grizzle, 1962; Laird 和 Ware, 1982; Singer, 1998); 在计量经济学文献中, 也称之为随机系数模型 (Rosenberg, 1973; Longford, 1993); 而在统计文献中则称之为协方差成分模型 (Dempster et al., 1981; Longford,

1987). 本书之所以采用分层线性模型这个术语, 是因为它表达了数据的一个重要结构特点, 这种数据使用范围广, 常见于增长性研究、机构效应和综合研究. 这个术语是由 Lindley 和 Smith(1972), Smith(1973) 引进的. 在这种前提下, Lindley 和 Smith 详细阐述了复杂的误差结构嵌入数据的普遍框架. 不过后续的研究曾经一度衰落, 因为这些模型的使用需要对不平衡的数据进行协方差成分的估计. 除了一些非常简单的问题之外, 在 20 世纪 70 年代早期没有一种全面的估计方法行得通. Dempster 等 (1977) 研究了 EM 的算法, 结果使之有了必要的突破: 一种概念上的可行和协方差成分估计的广泛应用方法. Dempster 等 (1981) 证明了这种分层数据结构方法的合理性. Laird 和 Ware(1982), Strenio, Weisberg 和 Bryk (1983) 把这种方法应用到增长性研究方面, 而 Mason 等 (1983) 则把它应用到多层结构横截面数据方面. 后来, 通过对最小二层重新反复地广泛使用和一种 Fisher 得分算法, 其他的多种协方差成分估计方法也就应运而生了 (Goldstein, 1986).

贝叶斯方法在这种情况下提供一种有意义的可选择方法. 标准误差比在 ML(极大似然) 下将趋向于更实际. 而且, 通过提供感兴趣的每个参数的后验分布, 贝叶斯方法提供有关研究问题的多种有兴趣的图表与数量证据的总结. 贝叶斯方法的出现并不新鲜. 比较新鲜的是较为方便的计算方法的出现, 尤其是在分层数据和模型的背景下. 这些新方法的进展是关于使用 Monte Carlo 方法来估计在先前被当作难处理的背景下后验分布的规则系统族. 这方法包括数据增广 (Tanner and Wong, 1987) 和 Gibbs 取样 (Gelfand et al., 1990; Gelfand and Smith, 1990).

3. 分位回归

关于传统的分层模型所用的统计分析方法主要是均值回归. 该方法有许多不足之处. Koenker 和 Bassett (1978) 提出了分位回归, 它可以看成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. 一个重要特殊的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最小化绝对误差的非对称的加权和. 简单地讲, 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, 另外, 它能估计出来的系数向量, 即边际效应, 对响应变量的离群观测值来说, 是稳健的; 给出在不同分位点上潜在的不同解, 这具有很有用的解释意义.

1.1.2　复杂数据概念

众所周知, 随着现代科学技术的飞速发展, 许多科学研究领域产生了多种复杂数据. 当然复杂数据的统计建模涵盖了许多当代统计分支, 推动了当代统计学理论方法的进步与发展, 并且其应用层面涉及生物信息学、流行病学和金融风险等, 意

义十分重大深远.

概念界定：本书所研究的复杂数据的明显特征之一是：高维、高频、多元、复杂的 "时空" 分层等性质. 数据分析方面的主要挑战来自：① 在高维空间中直接进行系统收缩变得非常困难; ② 一般高维函数的精确逼近很棘手; ③ 高维函数积分的实现变得不可能; ④ 对感兴趣的高维多元条件随机变量分布的全面刻画尚无先例可循; ⑤ 如果没有考虑普遍存在的复杂 "时空" 分层数据的特征, 常常使得传统的统计方法表现不佳, 甚至失效 ……

高维多元复杂数据的统计分析是目前全世界统计学界面临的最大挑战, 这无疑是当前统计学中的研究热点问题. 本书以此选题, 针对复杂数据相关问题开展研究.

1.1.3 传统模型

1. 线性分层分位回归模型

分层数据可以有很多层. 为了说明方便而且又不失一般性, 我们在本节里只考虑具有两层的数据, 有的书上又称为两水平数据. 其实, 所得到的基本结果是很容易推广到多层数据上. 假设有 $(\boldsymbol{X},\boldsymbol{W},\boldsymbol{Y})$ 的独立同分布观测值 $\{(\boldsymbol{X}_1,\boldsymbol{W}_1,\boldsymbol{Y}_1),\cdots,(\boldsymbol{X}_n,\boldsymbol{W}_n,\boldsymbol{Y}_n)\}$, 其中 $\boldsymbol{Y}_i$ 是实数响应变量的值, $\boldsymbol{X}_i$ 是已知的 $1\times d$ 维第一层预测值向量, $\boldsymbol{W}_i$ 是已知的 $d\times f$ 第二层预测矩阵, 满足第一层模型

$$\boldsymbol{Y}_i=\boldsymbol{X}_i\boldsymbol{\beta}_i+\varepsilon_i,\quad \varepsilon_i\sim N(0,\sigma^2),$$

其中 $\boldsymbol{\beta}_j$ 是未知的 $d\times 1$ 维第一层系数向量, ε_i 是 i.i.d. 不可观测随机效应变量, 假定它们与 $\boldsymbol{X}_i$ 独立并且服从均值为 0 方差为 σ^2 的正态分布.

在第二层模型上, 第一层模型中的系数成了输出结果:

$$\boldsymbol{\beta}_i=\boldsymbol{W}_i\gamma+\boldsymbol{u}_i,\quad \boldsymbol{u}_i\sim N(\boldsymbol{0},\boldsymbol{T}),$$

其中 $\boldsymbol{\gamma}$ 是 $f\times 1$ 维固定效应向量, $\boldsymbol{u}_i$ 是 $d\times 1$ 维第二层随机效应向量, 假定它们与 $\boldsymbol{W}_i$ 和 ε_i 独立并且服从均值向量为 0, 向量协方差阵为方阵 $\boldsymbol{T}_{d\times d}$ 的多元分布.

将第二层模型代入到第一层模型, 产生如下组合模型:

$$\boldsymbol{Y}_i=\boldsymbol{X}_i\boldsymbol{W}_i\gamma+\boldsymbol{X}_i\boldsymbol{u}_i+\varepsilon_i,\quad \varepsilon_i\sim N(0,\sigma^2),\quad \boldsymbol{u}_i\sim N(\boldsymbol{0},\boldsymbol{T}).$$

为了全面刻画给定预测变量 $(\boldsymbol{X},\boldsymbol{W})=(\boldsymbol{x},\boldsymbol{w})$ 的条件下响应变量 $\boldsymbol{Y}$ 的条件分布 $F(y|\boldsymbol{x},\boldsymbol{w})$, 我们来考虑 $\boldsymbol{Y}$ 的分位函数. 假定 $F(y\,|\boldsymbol{x},\boldsymbol{w})$ 是 y 的增函数并且在 $\boldsymbol{x}$ 和 $\boldsymbol{w}$ 连续, 那么在给定 $\boldsymbol{X}=\boldsymbol{x}$ 和 $\boldsymbol{W}=\boldsymbol{w}$ 的条件下, $\boldsymbol{Y}$ 的 τ 分位可定义为 $q_\tau(\boldsymbol{x},\boldsymbol{w})$, 它满足

$$q_\tau(\boldsymbol{x},\boldsymbol{w})=\inf\{t\in\mathbb{R}:F(t\,|\boldsymbol{x},\boldsymbol{w})\geqslant\tau\},\quad 0<\tau<1.$$

可以直接证明在上述组合模型之下, 有

$$q_\tau(\boldsymbol{x},\boldsymbol{w}) = \boldsymbol{x}\boldsymbol{w}\boldsymbol{\gamma} + (\boldsymbol{x}\boldsymbol{T}\boldsymbol{x}^{\mathrm{T}} + \sigma^2)^{1/2}\Phi^{-1}(\tau), \tag{1.1.1}$$

其中 $\Phi(\cdot)$ 是标准正态分布函数.

2. 半参数分层分位回归模型

为了简化陈述, 我们只考虑两层的分层模型. 结果很容易推广到三层或者更多层. 假设有 J 个单元, 每个单元有 n_j 个元素, 定义 $\boldsymbol{Y}_{ij}$ 是第 j 个单元的第 i 个个体. $\boldsymbol{Y}_{ij}$ 是一个实值随机变量, $\boldsymbol{X}_{ij}$ 是第一层 $d_1 \times 1$ 维非参数部分的解释向量, $\boldsymbol{Z}_{ij}$ 是 $d_2 \times 1$ 维参数部分解释向量.

第一层单元内部模型:

$$\boldsymbol{Y}_{ij} = m(\boldsymbol{X}_{ij}^{\mathrm{T}}) + \boldsymbol{Z}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_j + \varepsilon_{ij}, \quad j = 1, \cdots, J, \quad i = 1, \cdots, n_j,$$

其中 $m(\cdot)$ 是一个未知的函数. ε_{ij} 是不可观测的随机误差项, 其分布未知, $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \cdots, \beta_{jd_2})^{\mathrm{T}}$ 是一个 d_2 维的系数向量. 对于 $\boldsymbol{x}$ 点周围的 $\boldsymbol{X}_{ij}$, 有下面的局部线性展开:

$$m(\boldsymbol{X}_{ij}) \approx m(\boldsymbol{x}) + (\boldsymbol{X}_{ij} - \boldsymbol{x})^{\mathrm{T}}\nabla m(\boldsymbol{x}).$$

从而

$$\boldsymbol{Y}_{ij} \approx m(\boldsymbol{x}) + (\boldsymbol{X}_{ij} - \boldsymbol{x})^{\mathrm{T}}\nabla m(\boldsymbol{x}) + \boldsymbol{Z}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_j + \varepsilon_{ij} \triangleq \tilde{\boldsymbol{X}}_{ij}^{\mathrm{T}}\boldsymbol{\theta}_j + \varepsilon_{ij},$$

其中 $\tilde{\boldsymbol{X}}_{ij} = (1, (\boldsymbol{X}_{ij} - \boldsymbol{x})^{\mathrm{T}}, \boldsymbol{Z}_{ij}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\theta}_j(\boldsymbol{x}) = (m(\boldsymbol{x}), \nabla m(\boldsymbol{x})^{\mathrm{T}}, \boldsymbol{\beta}_j^{\mathrm{T}})^{\mathrm{T}}$.

第二层单元之间模型:

$$\boldsymbol{\theta}_j(\boldsymbol{x}) = \boldsymbol{W}_j\boldsymbol{\gamma}(\boldsymbol{x}) + \boldsymbol{U}_j,$$

其中 $\boldsymbol{W}_j$ 是 $(1 + d_1 + d_2) \times f$ 维的第二层的解释变量矩阵, $\boldsymbol{\gamma}(\boldsymbol{x})$ 是 $f \times 1$ 维的固定效应向量, $\boldsymbol{U}_j$ 是 $(1 + d_1 + d_2) \times 1$ 维的第二层的随机误差向量, 和 ε_{ij} 相互独立, 有均值向量 $E(\boldsymbol{U}_j) = 0$, 它与协方差阵 $\mathrm{cov}(\boldsymbol{U}_j) = \boldsymbol{T}$.

合并两层, 我们得到

$$\boldsymbol{Y}_{ij} = \tilde{\boldsymbol{X}}_{ij}^{\mathrm{T}}\boldsymbol{W}_j\boldsymbol{\gamma}(\boldsymbol{x}) + \tilde{\boldsymbol{X}}_{ij}^{\mathrm{T}}\boldsymbol{U}_j + \varepsilon_{ij}.$$

令 $\xi_{ij} = \tilde{\boldsymbol{X}}_{ij}^{\mathrm{T}}\boldsymbol{U}_j + \varepsilon_{ij}$, 其中 $\xi_{ij} \sim G_{ij}$. 计 $F(y)$ 为 $\boldsymbol{Y}$ 的分布函数. 为了得到给定条件条件 $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) = (\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w})$ 下响应变量 $\boldsymbol{Y}$ 的分布 $F(y|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w})$, 我们考虑 $\boldsymbol{Y}$ 的分位函数. 假定 $F(y|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w})$ 是 y 的增函数, 且在 $\boldsymbol{x}$ 和 $\boldsymbol{w}$ 处连续, 于是在条件 $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) = (\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w})$ 下 $\boldsymbol{Y}$ 的 τ 阶分位数定义为 $q_\tau(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w})$. 可以直接证明在合并后的模型之下, 有

$$q_\tau(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}) = \inf\{t \in \mathbb{R} : F(t|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}) \geqslant \tau\}, \quad 0 < \tau < 1.$$

此为试读, 需要完整PDF请访问: www.ertongbook.com