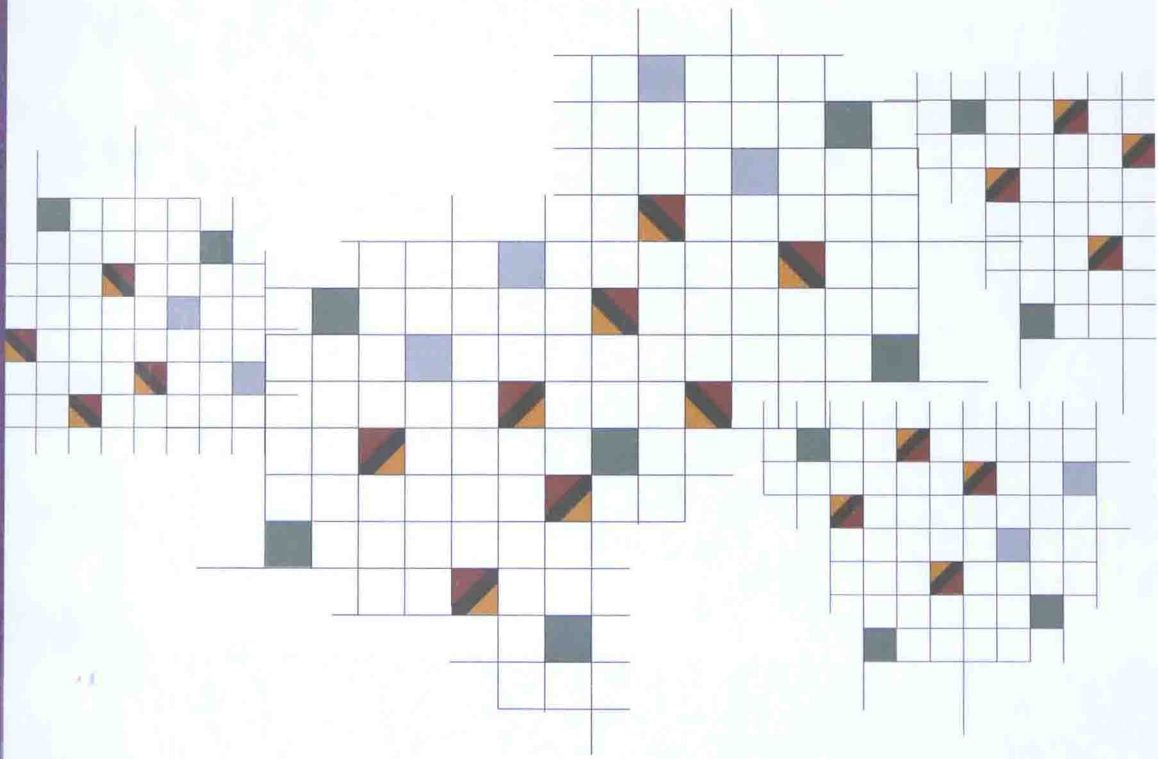
田茂再 著



复杂数据统计推断
理论、方法及应用

科学出版社

复杂数据统计推断理论、方法及应用

田茂再　著

科学出版社

北京

内 容 简 介

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，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，并且其应用层面几乎涉及各领域．具有复杂分层结构的数据在现实生活中很普遍．能完全剖析这类数据，发掘该类数据表象下的潜在规律性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前沿知识，侧重于系统的理论与算法介绍．内容主要涉及线性分位回归、非参数分位回归、适应性分位回归、可加性分位回归、变系数分位回归、单指数分位回归、分位自回归、复合分位回归、高维分位回归以及贝叶斯分位回归、分层样条分位回归、分层线性分位回归、分层半参数分位回归、复合分层线性分位回归以及复合分层半参数分位回归，等等．

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

复杂数据统计推断理论、方法及应用/田茂再著. —北京：科学出版社，2014

ISBN 978-7-03-040499-2

Ⅰ.①复… Ⅱ.①田… Ⅲ.①统计学 Ⅳ.①C8

中国版本图书馆 CIP 数据核字(2014)第 082163 号

责任编辑：刘凤娟／责任校对：彭 涛

责任印制：肖 兴／封面设计：王 浩

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

北京佳信达欣艺术印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2014 年 5 月第 一 版 开本：720×1000 1/16

2014 年 5 月第一次印刷 印张：21 1/2

字数：410 000

定价：118.00 元

(如有印装质量问题，我社负责调换)

前　言

随着科学技术的飞速发展, 许多科学研究领域产生了多种多样的海量超高维复杂数据. 这些领域包括基因学、天文学、宇宙学、流行病学、经济学、金融学、功能性磁共振成像以及图像处理等. 面对这些高速增长的复杂超高维海量数据的挑战, 要求各个领域的科学家具有快速提取所需信息的能力. 因此, 就统计学自身而言, 通过对这些复杂数据的统计推断, 研发出强有力的统计科研工具, 显然会给统计界带来切实的利益; 将有利于统计学科理论和方法在更广阔的天地中长足发展, 有利于促进对自然和科学的深度理解. 因此, 我们不难发现眼下学术造诣深厚的国际统计学大家已经将他们的研究兴趣转移到了复杂数据工程上来. 对复杂数据开展深入系统的创新性研究, 引进新思想, 研发新工具, 形成新理论, 从而推动其他重要领域和科学前沿取得突破. 其实, 随着大量产生于当今科学的复杂数据不停地快速增长, 从基因组到自然科学领域, 统计学家一直在积极参与跨学科领域的科学研究. 从统计学的发展史可以看出, 各门具体科学领域产生的复杂数据挑战越多, 统计学家面临的机遇也就越多, 相关的统计学理论和方法也得到了空前的发展. 反过来, 这些理论和方法也推动着许多重要领域或科学前沿取得突破. 本书致力于介绍复杂数据分析前沿的统计理论和方法.

关于复杂数据概念界定, 有很多不同的本书所研究的复杂数据的明显特征之一是具有高维、高频、多元或者复杂的 "时空" 分层结构等. 这类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搜索变得非常困难; 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很难, 甚至不可能; 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. 所以, 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. 基于作者前期工作的微薄积累, 本书针对复杂数据相关问题开展研究. 粗略地说, 本书即将着手分析的数据可分为 6 大类 (不严格地说): 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); ⑥ 有序分类数据 (ordinal categorical data). 重点解决下列当代统计学前沿问题: 具有 "复杂时空" 等结构的数据建模, 建立一套完善的能刻画该类型数据各层面特征的几大类 "分层分位回归模型" 的理论与方法, 并付诸实际应用.

本书的目的就是为读者提供一些复杂数据分析的知识, 侧重于理论与方法的系统性论述.

由于作者水平有限, 疏漏之处在所难免, 甚望批评指正!

本书得到下面基金的资助: 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) 以及甘肃省教育厅“飞天学者”特聘教授计划项目. 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科学研究中心的大力支持.

田茂再

2013 年 6 月

目　录

第一部分　分位回归

第二部分　分层分位回归

第一部分

分 位 回 归

第 1 章　分位回归引言

1.1　概　　述

分位回归由 Koenker 和 Bassett (1978) 提出, 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.

1.1.1　分位数定义

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 (无条件地) 定义为

$$Q_\tau(Y) = \operatorname{Arg}\inf\{y \in \mathbb{R}; F(y) \geqslant \tau\}, \quad 0 < \tau < 1.$$

若将分布函数 $F(x)$ 的逆定义为 $F_Y^{-1}(\tau) = \inf\{y \in \mathbb{R}; F(y) \geqslant \tau\}$, 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; 中位数是分位数的一个最熟知的例子. 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 $P(Y \leqslant m) = P(Y \geqslant m) = \dfrac{1}{2}$. 特别地, 对于一个连续型随机变量, m 是等式 $F(m) = \dfrac{1}{2}$ 的一个解, 其中 $F(y) = P(Y \leqslant y)$ 为累积分布函数. 作为中位数的用法的一个例子, 我们考虑工资的分布. 由于只有少数的人赚取巨额的工资, 所以这个分布是典型右偏的. 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.

更一般地, 25%样本分位数可以被定义为将数据分割成四分之一和四分之三两部分的值, 反过来, 可以定义 75%样本分位数. 相应地, 连续情形中总体的下四分之一分位数和上四分之三分位数各自为等式 $F(y) = \dfrac{1}{4}$ 和 $F(y) = \dfrac{3}{4}$ 的解. 一般地, 对于一个比例 $\tau(0 < \tau < 1)$, 在连续情形中, F 的 $100\tau\%$分位数 (等价地, 第 100τ 的百分位数) 是 $F(y) = \tau$ 的解 y, 假定解是唯一的.

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 $F(x)$ 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 $F_n^{-1}(\tau) = X_{([n\tau])}$, 其中符号 $[\cdot]$ 表示 $\cdot$ 的取整.

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 (如 80%) 比例的学生好, 同时比 $1-\tau$ (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, 而另一半则表现比中位数差. 类似地, 四分位数将总体分为四段, 在每一段中所占相应于总体的比例是相同的. 五分位数将总体分为五部分; 十分位数则将总体分为十部分. 在一般情况下, 分位数又称为百分位数, 或者有时候又称作分位数. 分位回归由 Koenker 和 Bassett (1978) 提出, 看在估计条件分位函数模型, 模型中响应变量条件分布的分位数表示为观察到的协变量的函数.

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

$$Q_\tau(Y|X = x) = \operatorname{Arg}\inf\{y \in \mathbb{R}; F(y|x) \geqslant \tau\}, \quad 0 < \tau < 1.$$

1.1.2　分位回归

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, 而分位回归测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. 下面这幅图 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趋势, 并非成直线趋势. 同时也注意到中位数回归曲线与均值回归曲线接近.

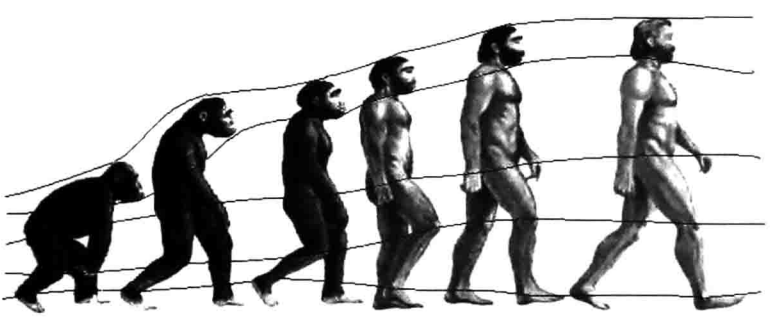


图 1-1　人类进化曲线

下面从模型角度来讲. 假定有样本序列 $\{(X_i, Y_i), i = 1, \cdots, n\}$ 满足下列回归模型:

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

其中 X_i, $i = 1, \cdots, n$ 为固定设计点. 假定误差项 ε_i, $i = 1, \cdots, n$ 为独立同分布的序列, 且分布情况未知, 则响应变量 Y 的 τ 阶条件分位 $m_\tau(x)$ 满足 $\tau = P(Y \leqslant m_\tau(X)|X = x)$. 经过简单地计算, 亦可等价地定义为

$$m_\tau(x) = \operatorname{Arg}\min_{\theta \in \mathbb{R}} \mathbb{E}\{\rho_\tau(Y - \theta)|X = x\}, \tag{1.1}$$

其中 $\rho_\tau(u) = u\{\tau I(u \geqslant 0) - (1 - \tau)I(u < 0)\}$ 为检验函数, $I(\cdot)$ 为示性函数, 不包含示性函数的检验函数的写法为

$$\rho_\tau(u) = \begin{cases} \tau u, & u \geqslant 0, \\ (\tau - 1)u, & u < 0. \end{cases}$$

检验函数是损失函数的一种, 直观得知检验函数均为正值, 且分位数 τ 会影响检验函数的值, 其中 τ 为我们所感兴趣的分位数.

1. 损失函数与风险函数

所谓损失函数 (loss function), 就是定量描述决策损失大小程度的函数, 下面的图 1-2 中展示了三种不同形式的损失函数, 其中包括平方损失、绝对值损失和检验函数.

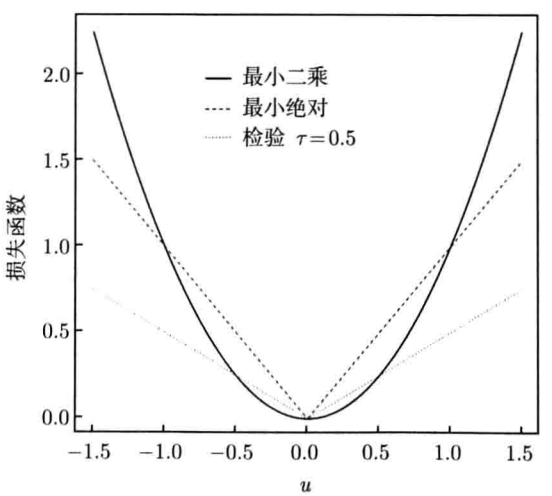


图 1-2　三种类型的损失函数

(1) 平方损失函数 (the square loss function)

平方损失函数也称为二次损失函数, 通常将平方损失函数定义为 $l(y, f(x)) = (y - f(x))^2$, 均值回归估计在平方损失在意义下的理论基础为 "残差平方和最小", 即最小二乘估计 (least square estimate) 的基本思路. 然而, 平方损失函数的一个明显缺陷为异常点 (outlier point) 存在的情形, 远离目标函数的孤立点往往对目标函数的影响极大, 异常点存在的情形会急剧扩增或缩减 $l(\cdot)$ 的预测值, 因此在通常经济统计、社会统计数据分析中, 为了正常使用最小二乘方法而通常首先将数据净化 (filtered data), 即去掉异常点.

(2) 绝对值损失函数 (the absolute value loss function)

同平方损失类似, 绝对值损失也是一种衡量对统计决策造成损失的大小的一种方法, 由于二次损失中存在的平方效应, 绝对值损失的优点就显而易见, 其优点就在于其波动程度比平方损失小, 因此在某些情况下绝对值损失函数有其特定优势.

(3) 检验函数 (check function)

分位回归领域有其专门的损失函数, 我们称之为检验函数, 该函数与被估计模型的 τ 分位相关. 因此函数中多了分位数这一变元, 在通常分位数回归中, 分位数通常是给定的, 即做人们感兴趣的某一分位上的回归曲线. 当然, 近年来也有针对分位数选择的新方法提出. 例如, 检验函数的形式有多种, 但是万变不离其根本, 即下面的表达式

$$\rho_\tau(u) = (\tau I_{[u\geqslant 0]} + (1-\tau) I_{[u<0]})|u| = (\tau - I_{[u<0]})u.$$

由检验函数的定义以及图 1-2 都可以看出检验函数在原点不连续, 即有在原点不可导.

2. 分位数的样本实现

分位数似乎与样本观测量的排序和分类过程密不可分. 因此, 令人惊讶的是, 可以通过另一个简单的权宜之计将分位数定义为一个最优化问题. 就像能够定义样本均值为最小化残差平方和问题的解一样, 可以定义中位数为最小化绝对残差和的解. 那对于其他的分位数呢? 如果对称的绝对值函数产生中位数, 可以简单地尝试将绝对值倾斜以便得到一个非对称加权从而产生了其他的分位数. 这个"弹球游戏规则"建议求解如下问题:

$$\min_{\xi\in\mathbb{R}} \sum \rho_\tau(y_i - \xi). \tag{1.2}$$

为了看到这个问题会导致样本分位数作为它的解, 就必须要计算目标函数关于 ξ 的方向导数, 分别是从左和从右.

我们已经成功地定义无条件分位数为一个最优化问题, 那么以一个相似的方式定义条件分位数就很简单了. 最小二乘回归为如何进行提供了一个模型. 如果对于