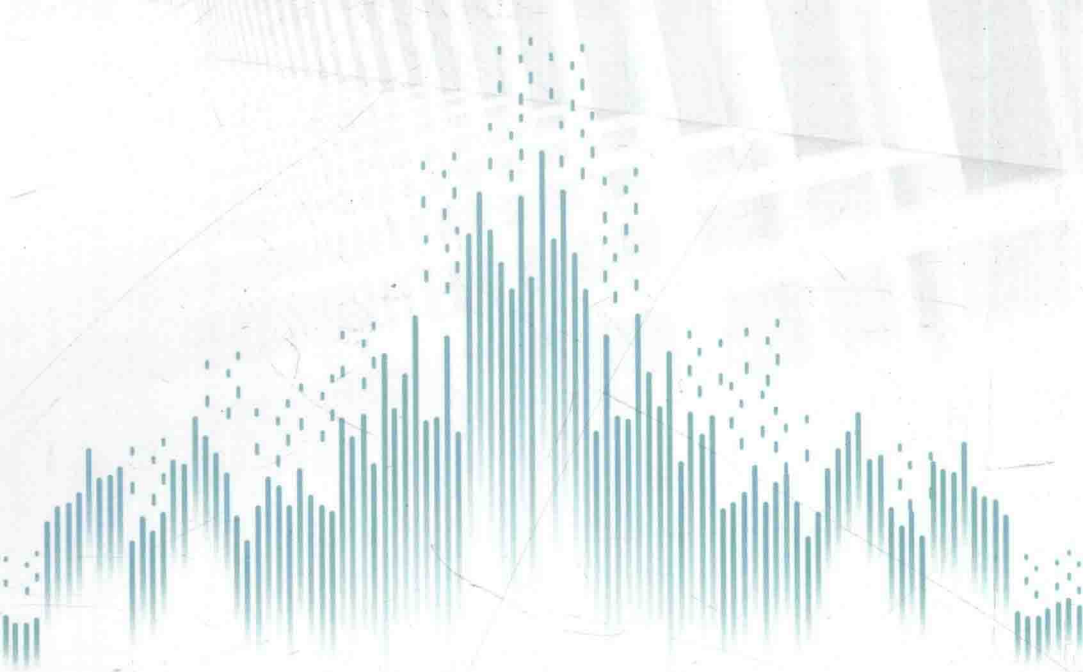

基于大规模数据的分位数回归模型及应用

JIYU DAGUIMO SHUJU DE FENWEISHU HUIGUI MOXING JI YINGYONG

蔡超／著

中国财经出版传媒集团
经济科学出版社
Economic Science Press

基于大规模数据的分位数回归模型及应用

JIYU DAGUIMO SHUJU DE
FENWEISHU HUIGUI MOXING JI YINGYONG

蔡超 / 著

中国财经出版传媒集团
经济科学出版社
Economic Science Press

图书在版编目（CIP）数据

基于大规模数据的分位数回归模型及应用/蔡超著.
—北京：经济科学出版社，2018.7
ISBN 978-7-5141-9238-4

Ⅰ.①基… Ⅱ.①蔡… Ⅲ.①自回归模型
Ⅳ.①O212.1

中国版本图书馆 CIP 数据核字（2018）第 080721 号

责任编辑：李 雪 程辛宁
责任校对：隗立娜
责任印制：邱 天

基于大规模数据的分位数回归模型及应用
蔡 超 著
经济科学出版社出版、发行 新华书店经销
社址：北京市海淀区阜成路甲 28 号 邮编：100142
总编部电话：010-88191217 发行部电话：010-88191522
网址：www.esp.com.cn
电子邮件：esp@esp.com.cn
天猫网店：经济科学出版社旗舰店
网址：http://jjkxcbs.tmall.com
固安华明印业有限公司印装
710×1000 16 开 12 印张 160000 字
2018 年 7 月第 1 版 2018 年 7 月第 1 次印刷
ISBN 978-7-5141-9238-4 定价：48.00 元

- 国家自然科学基金项目“基于高维非线性广义分位数回归的系统性金融风险计量”(编号:71671056)
- 山东省社会科学规划项目“大规模数据的惩罚分位数回归方法研究”(编号:18DTJJ01)
- 山东工商学院博士启动基金项目“流数据分位数回归模型及应用研究”(编号:BS201815)

前　言

大数据时代，随着数据生成、收集与存储技术的发展，以大样本与高维为典型特征的大规模数据将会大量涌现。这为探索客观规律带来了机遇，也为统计分析带来了挑战。在统计方法中，分位数回归常用来反映解释变量对响应变量整个条件分布的异质影响，能够细致刻画响应变量的尾部行为，是探索客观规律的重要手段与方法之一。常用的统计软件都可进行分位数回归，但受到计算内存和运行时间的限制，大规模数据分位数回归往往难以奏效。因此，在大数据背景下，研究大规模数据分位数回归方法，解决其建模过程中的技术难题，对于推广应用、揭示经济和社会的复杂模式等，具有重要的理论意义和实践价值。

本书选取“基于大规模数据的分位数回归模型及应用”这一研究主题，综合应用统计学和计量经济学等学科知识，采取理论分析、数值模拟和应用研究相结合的范式，将经典的分位数回归模型从中小规模数据扩展到大规模数据。

本书共分为7章，第1章为绪论。主要介绍了本书选题的研究背景、研究意义及研究现状等。第2章为经典分位数回归模型。主要介绍了线性分位数回归模型、非线性分位数回归模型和惩罚分位数回归模型等。第3章为基于稀疏指数转移方法的大样本数据分位数回归及应用。主要提出了基于稀疏指数转移方法的大样本数据分位数回归模

型，通过数值模拟，研究其估计效果、预测能力以及运行时间，并将其应用于中国股票市场，研究股票收益与指令不均衡之间的异质性关系。第4章为基于随机抽样算法的大规模数据套索惩罚分位数回归及应用。主要提出了基于随机抽样算法方法的大规模数据套索惩罚分位数回归模型，通过数值模拟，研究其估计效果、变量选择能力、预测能力以及运行时间，并将其应用于选取美国温室气体监测数据，实证检验其估计和预测能力、变量选择能力。第5章为基于分块估计方法的大样本数据分位数回归及应用。主要提出了基于分块估计方法的大样本数据分位数回归模型，通过数值模拟，研究其估计效果、渐近正态性和预测能力，并将其应用于中国劳动力市场，研究教育和工作经验与收入之间的异质性关系。第6章为基于分块估计方法的大规模数据套索惩罚分位数回归及应用。主要提出了基于分块估计方法的大规模数据套索惩罚分位数回归模型，通过数值模拟，研究其估计效果、变量选择能力、渐近正态性和预测能力，并将其应用于美国温室气体监测数据，实证检验其估计和预测能力、变量选择能力。第7章为总结与展望。

本书的部分内容是国家自然科学基金项目“基于高维非线性广义分位数回归的系统性金融风险计量”（编号：71671056）、山东省社会科学规划项目“大规模数据的惩罚分位数回归方法研究”（编号：18DTJJ01）、山东工商学院博士启动基金项目“流数据分位数回归模型及应用研究”（编号：BS201815）的阶段性成果。

目录

第1章

绪　论

本章主要介绍本书选题的研究背景及研究意义，综述了大规模数据分析技术与分位数回归方法的国内外研究现状，并从中发掘出有待进一步探究的问题。最后，介绍了本书的结构安排与主要创新工作。

1.1　研究背景和意义

1.1.1　研究背景

大数据指的是具有撷取信息功能的结构化、半结构化和非结构化的规模巨大数据。大数据具有体量巨大（volume）、种类繁多（variety）、流动速度快（velocity）、价值密度低（value）这“4V”特征。大数据时代，随着数据生成、收集与存储技术的发展，数据规模呈现爆炸式增长（涂新莉等，2014）。例如，对经济个体的调查数据，不

再局限于性别、年龄、收入、教育等传统数据，而是通过先进的计算机技术与手机终端，实现对经济个体的实时监测，获取个人电子商务消费、网页浏览历史、社交网络交流等数据（耿直，2014）；对金融市场的调查数据，不仅包含股票指数、交易量等数据，而且通过先进的计算机技术与手机终端，获取微博数据、网络媒体报道信息等（杨莎等，2015；王春峰等，2016）。大数据使得人们能够通过数据观察分析经济、社会的客观现象，例如，迈尔－舍恩伯格（Mayer－Schnberger，2013）指出推特和脸谱网通过用户的社交网络图获得用户的喜好。因此，大数据的广泛应用，对于探索与揭示经济和社会的运行模式与客观规律，以及推进经济和社会的持续健康发展，具有重要的理论意义和实践价值。

涂兰敬（2011）和陶雪娇等（2013）指出大规模数据又称海量数据，主要包括结构化和半结构化的数据，是以大样本与高维为典型特征的数据集。大数据包含了大规模数据的含义，而且在内容上超越了大规模数据。简而言之，大规模数据是大数据的组成部分。之前，由于数据量相对不足，对数据的研究是“由薄变厚”，把“小”数据变“大”。而在大数据时代，是要把数据“由厚变薄”，将数据去冗分类、去粗存精。因此，大规模数据为准确揭示现实世界的客观规律带来了机遇，也为统计分析带来了挑战，需要在方法论上有所突破。在统计分析中，回归分析常被用以讨论解释变量对响应变量的影响，是探索客观规律的重要手段与方法之一。最常用的工具是均值回归（mean regression）方法，主要目的在于通过解释变量来估计和预测响应变量的条件均值，从而揭示解释变量和响应变量间的真实关系。建立在古典假定基础上的均值回归，虽然具有线性性、无偏性和有效性等优良特性，但仍存在局限：一是均值回归仅能度量响应变量的条件均值，而无法给出响应变量的整个分布特征；二是均值回归一般假设

误差项独立同分布，且服从正态分布。然而，现实世界中，古典假定的条件往往难以得到满足。康克等（Koenker et al.，1978）提出的分位数回归，能够利用解释变量得到响应变量条件分布在多个分位点的分位数函数。与传统的均值回归相比，它既可以描述自变量对于响应变量的变化范围以及条件分布形状的影响，又能够不受异方差的限制，得到更加稳健的结果。经过近 40 年的发展，分位数回归广泛应用于经济管理学领域和社会学领域等。

常用的统计软件都可进行分位数回归，但对大规模数据进行分位数回归建模时，将面临两个主要的困难：第一，计算内存限制。目前，大多数统计方法只适用于全部数据放在单个计算机内存的环境。然而，大规模数据体量巨大，当其读入时，就已经占据了大量内存，再对其进行计算时，则计算内存严重不足。这一局限已经被大量理论与实践所证实，例如，科恩（Cohen，2009）、张延松等（2011）和何清等（2014）都指出，计算机内存越来越不能负担大规模数据的存储和计算。第二，运行时间约束。处理大规模数据往往需要大量时间，不能及时给出有效结果。例如，李仲达（2015）指出，对高维数据进行线性回归时，其嵌套的子模型是变量维数的两倍，为了获得最优子模型，需要对子模型逐个估计，当变量维数较大时，运行时间无疑是巨大的；张素香等（2015）指出，当数据量巨大时，传统局部加权线性回归预测方法需要为每个测试点寻找近邻，运算量很大，单机运算的时间可能会达到几个小时或几天。目前，大规模数据的统计建模已经成为众多学者追踪和关注的热点问题。克拉克森（Clarkson，2005）提出的随机映射和随机抽样算法和樊采虹等（Fan Tsai - Hung et al.，2007）、李润泽等（Li Runze et al.，2013）提出的分块估计方法，都为大规模数据的分析提供了一个基本工具，证实了在减少占用的计算内存和降低运行时间的同时，也能够获得一个精确统计

推断结果。

因此，在大数据背景下，研究大规模数据分位数回归方法，解决其建模过程中的技术难题，对于推广应用、揭示经济和社会的复杂模式等，具有重要的理论意义和实践价值。

1.1.2 研究意义

目前，分位数回归的研究主要集中在两个方面：第一，在理论建模方面，对经典分位数回归模型进行有意义的扩展，开发出新的分位数回归模型与方法，丰富分位数回归的理论研究内容。第二，在应用研究方面，将分位数回归理论应用于更多学科领域，拓展分位数回归的应用范畴。尽管分位数回归在上述方面都已取得了丰硕成果，但上述方面多集中在中小规模数据，对于大规模数据分位数回归仍有待于进一步探讨。鉴于此，本书选取“基于大规模数据的分位数回归模型及应用”这一研究议题，在大规模数据背景下，对传统分位数回归模型与方法进行深入研究与扩展改进，分别探讨基于稀疏指数转移方法的大样本数据分位数回归、基于随机抽样算法的大规模数据套索惩罚分位数回归、基于分块估计方法的大样本数据分位数回归和基于分块估计方法的套索惩罚大规模数据分位数回归等建模技术，并将其应用于实际问题解决。本书的研究意义在于：

（1）从理论角度。第一，在模型设定上，上述方法都是对经典分位数回归方法的在大规模数据背景下的有意义扩展，都具备独特优点：上述方法不仅能够取得精确的估计值，而且能够减少计算内存需求和降低运行时间，即可在普通计算机上计算并及时给出有效结果。第二，在建模技术上，深入研究大规模数据分位数回归方法的建模步骤以及参数估计值的渐近性质，为全面深入地掌握各方法的统计特征

与优良性能，提供了相应的理论依据。

（2）从应用角度。经济管理领域往往涉及大规模数据，选取经济管理领域的热点问题，在大规模数据分位数回归框架下开展相关主题研究。第一，将大规模数据分位数回归应用于股票收益率与指令不均衡之间关系研究，细致地揭示指令不均衡对股票收益率整个条件分布的影响，并且实现股票收益率条件密度的准确预测，这一工作有助于投资者了解和掌握股市指令不均衡变化所预示的股票未来收益率的变动规律，指导其针对不同股票制定相应的风险防范措施和投资策略。第二，将大规模数据分位数回归应用于中国劳动力市场，研究教育和工作经验对收入整个条件分布的影响，这一工作有助于管理者了解和掌握劳动力市场收入的变动规律，指导其针对居民收入差距制定相应的措施。

1.2 国内外研究现状

1.2.1 大规模数据分析技术

近年来，随机映射和随机抽样算法被广泛应用于大规模数据回归分析中，其主要思想为：首先，构造一个抽样矩阵；其次，运用抽样矩阵从大规模数据中抽取一个子样本；最后，运用抽取的子样本进行回归分析。这种方法在规定的误差范围内，通过抽样方法来实现大规模数据回归分析，既能够得到一个精确的参数估计值，又可以减少占用的计算内存和降低运行时间。克拉克森（Clarkson，2005）、索勒等（Sohler et al.，2011）将随机映射和随机抽样算法应用于大规模

数据的 L_1 回归；德瑞内斯等（Drineas et al.，2006）将随机抽样算法应用于大规模数据的 L_2 回归；对于大规模数据 L_p 回归问题的求解，克拉克森等（Clarkson et al.，2013）运用了快速柯西转移（fast cauchy transform）方法，孟祥瑞等（Meng Xiangrui et al.，2013）运用了低扭曲子空间嵌入抽样（low-distortion subspace embedding sampling）算法，伍德拉夫等（Woodruff et al.，2014）运用了稀疏指数转移（sparse exponential transform）方法，并指出稀疏指数转移方法比快速柯西转移方法及低扭曲子空间嵌入抽样算法的运行时间更短。杨继言等（Yang Jiyan et al.，2013，2014）将快速椭圆取整（fast ellipsoid rounding）和快速柯西转移方法应用于大规模数据分位数回归，结果表明其估计结果与全样本的估计结果非常接近，而且减少了占用的计算内存和降低了运行时间。

目前，将随机映射和随机抽样算法应用于大规模数据的研究取得了一定成果，然而，随机映射和随机抽样算法有两个明显的缺陷：第一，要求数据能够储存在同一个内存中来进行随机抽样，当数据太大超出内存储存限制时，随机抽样将不能执行，导致方法失效；第二，随机抽样只利用了部分数据，没有充分发挥完整数据信息的优势。樊采虹等（2007）和李润泽等（2013）提出了分块估计方法，其基本思想是：首先，将大规模数据划分成若干个块；其次，对每一个块进行统计分析；最后，将每个块的参数估计值进行简单平均，作为全样本估计值的近似结果。分块估计方法能够很好地解决随机映射和随机抽样算法中的上述两个缺陷：一方面，将数据分块处理，降低了内存需求；另一方面，充分利用完整数据信息，能够获得更精确的结果。例如，樊采虹等（2007）运用分块估计方法解决均值回归问题，并在理论上证明了其渐近性质，最后使用模拟数据和信用卡的实际数据进行实证研究，结果表明估计结果与全样本回归结果基本一致。樊采

虹等（2007）将分块估计方法应用于逐步回归，并使用模拟数据和美国1994～1995年人口普查数据进行实证研究，结果表明估计结果与全样本回归结果相比，没有发生较大偏差。陈雪莹等（Chen Xueying et al.，2014）将分块估计方法应用于带有套索（least absolute shrinkage and selection Operator，LASSO）、平滑截断绝对偏差（smoothly clipped absolute deviation，SCAD）和极大极小凹（minimax concave penalyt，MCP）等惩罚的均值回归，并证明了其优良性质，最后使用模拟数据进行实证研究，结果表明估计结果具有较高精度，并且实现了变量选择。赵天琪等（Zhao Tianqi et al.，2014，2016）将分块估计方法应用于半参数回归和复合分位数回归，并证明了其优良性质。常香玉等（Chang Xiangyu et al.，2017）将分块估计方法应用于局部平均回归，研究发现基于分块估计的局部平均回归能够达到最优学习速度。

目前，关于随机映射和随机抽样算法以及分块估计方法的研究虽然取得了一定成果，但主要工作集中在均值回归等方面，尚有大量问题有待于探讨。例如，将随机映射和随机抽样算法或分块估计方法应用于分位数回归和惩罚分位数回归模型，并证明其理论性质等，本书对此进一步开展研究。

1.2.2 分位数回归方法

回归分析有着悠久的历史，建立在古典假定基础上的均值回归方法应用最为广泛，主要用于描述解释变量对响应变量条件均值的影响，具有线性性、无偏性、有效性等优良性质。但在实际问题中，古典假定往往不被满足，且当数据散布较大时，均值回归只能得到一条回归曲线，往往难以具有代表性。

康克等（1978）提出的分位数回归方法能够很好地弥补均值回归的上述缺点，而且具有更好的稳健性。分位数回归采用加权残差绝对值之和的方法估计参数，与均值回归相比，有如下优点：第一，可以细致地刻画解释变量对响应变量整个条件分布的影响；第二，对模型的随机扰动项无须做分布假定，提升了模型构建的稳健性；第三，对响应变量具有单调变换性；第四，参数估计在大样本理论下具有渐近优良性。①

近40年来，在理论模型和应用研究等方面，众多学者对分位数回归开展了广泛深入的研究。第一，在模型计算方面，主要方法包括单纯形算法（康克等，1987）、内点算法（Portnoy et al.，1997）和平滑算法（Chen Colin，2004，2007）等。第二，在模型检验方面，主要包括用以诊断回归系数与回归方程显著性的沃尔德检验（Wald）和似然比检验、拟合优度检验（康克等，1999）和回归系数同质性检验和对称性检验（康克等，1982）。第三，在模型扩展方面，随着计量技术的发展，分位数回归的建模框架不断得以丰富，例如，康克等（2006）提出的分位数自回归（QAR）模型、肖志杰（Xiao Zhijie，2009）提出的分位数协整模型和许启发等（2011）提出的分位数局部调整模型。第四，在模型预测方面，分位数回归能够细致刻画响应变量在不同分位点处的变动规律，不仅能够提供点预测（包含：均值、中位数和众数预测）（Engle et al.，2004；Gneiting，2011；陈磊等，2012）与区间预测（Granger，1989；Taylor，2007；蔡宗武等，2012），而且能够提供条件密度预测（许启发等，2011；何耀耀等，2013；阮素梅等，2015），以反映响应变量条件分布的完整信息。

① 更为详细的介绍，参见康克（2005）。