2024 届本科生学士学位论文　　学校代码：10269


華東師範大學

East China Normal University

# 本科生毕业论文

# 带辅助信息的分位数回归算法

# An Algorithm for Quantile Regression with Summary Information


| | |
|---|---|
| 姓　　名: | 吴昊 |
| 学　　号: | 10190720401 |
| 学　　院: | 统计学院 |
| 专　　业: | 统计学 |
| 指导教师: | 周勇 |
| 职　　称: | 教授 |
| 完成时间: | 2024 年 5 月 |

# 华东师范大学学位论文诚信承诺

本毕业论文是本人在导师指导下独立完成的，内容真实、可靠。本人在撰写毕业论文过程中不存在请人代写、抄袭或者剽窃他人作品、伪造或者篡改数据以及其他学位论文作假行为。

本人清楚知道学位论文作假行为将会导致行为人受到不授予/撤销学位、开除学籍等处理（处分）决定。本人如果被查证在撰写本毕业论文过程中存在学位论文作假行为，愿意接受学校依法作出的处理（处分）决定。

承诺人签名：_____　　日期：_____ 年 __ 月 __ 日

# 华东师范大学学位论文使用授权说明

本论文的研究成果归华东师范大学所有，本论文的研究内容不得以其它单位的名义发表。本学位论文作者和指导教师完全了解华东师范大学有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅；本人授权华东师范大学可以将论文的全部或部分内容编入有关数据库进行检索、交流，可以采用影印、缩印或其他复制手段保存论文和汇编本学位论文。

保密的毕业论文（设计）在解密后应遵守此规定。

作者签名：_____　导师签名：_____日期：_____ 年 __ 月 __ 日

# 目录

# 摘　要

现代统计学发展的一百多年来，统计学家们在实际问题的驱动下，由均值回归发展衍生出了一套回归理论和方法。然而，随着社会的进步以及人们看待问题的角度不断深入，单纯的均值回归已经不能满足人们对于分析工具的需要。人们并不总是只关心解释变量对响应变量均值的影响，有时可能关心对哪个等级的响应变量有什么样的影响。在这样实际问题的驱动下，统计学家发展出了分位数回归模型，通过对响应变量条件分位数的建模，为人们分析现实问题提供了更多更灵活的角度。分位数回归自提出半个世纪以来，其应用场景越来越多，相比于均值回归，其优越性在很多实际问题中已经得到了验证，已然成为近年来统计学研究的一大热点。

21 世纪以来，计算机的普及给人们的生活带来了大量的便利，推动了社会进步和生产力的发展，近十年来算力爆发式的增长更是推动社会进入了大数据时代，每天都有亿万级别的数据被产生、传输和储存，在大数据时代，对于感兴趣的问题获得与它相关的信息，也就是辅助信息是相当容易的。因此，在统计学界，如何引入辅助信息提高参数估计效率或是预测精度，也逐渐成为备受关注的热点问题之一。

然而，面对海量的辅助信息，就不得不考虑引入的辅助信息的随机性，在分位数回归中引入辅助信息已被研究者以各种角度探索过，然而，在分位数回归中引入辅助信息并且考虑辅助信息的随机性这一领域还鲜有研究。因此本文提出在分位数回归模型中引入辅助信息，并考虑辅助信息的随机性，设计算法求解分位数回归系数的估计，通过引入辅助信息来提高估计效率，并通过考虑辅助信息的随机性给模型带来更多的信息，以此提升参数估计的稳健性。

在本文第三章，针对考虑辅助信息随机性的分位数回归模型提出了核光滑的方法，将非光滑的分位数损失变成光滑函数，从而设计了一套便于计算的线性迭代算法求解回归系数的估计。并且由于分位数回归的系数估计问题实际上是一种特殊的 U 过程的参数估计问题，所以本文使用针对目标函数进行扰动重抽样的方法估计了回归系数的方差，这是一种广义的 Bootstrap 方法。

随后本文针对提出的算法进行了模拟实验，实验表明，带随机性辅助信息的模型对于参数估计的偏差相比不带辅助信息的分位数回归模型更小，方差相近，并且经验覆盖率更接近理论值，这意味着本文提出的算法通过引入辅助信息并考虑其随机性，确实提高了参数估计的效率，并且回归系数相比不带辅助信息的系数估计具有更好的渐进性质，进一步说明了算法的有效性。

本文旨在通过在分位数回归中引入辅助信息并考虑随机性，希望能为如何提高参数

估计的效率以及引入辅助信息的问题提供新的视角，并启发未来的研究者在此领域进行更深入的探索和创新。

**关键词:** 分位数回归，辅助信息，扰动重抽样，Bootstrap，U 过程，迭代算法，估计方程

# Abstract

In the more than one hundred years of the development of modern statistics, statisticians have developed a set of regression theories and methods derived from mean reversion driven by practical problems. However, with the progress of society and the deepening of people's perspectives on problems, simple mean regression can no longer satisfy people's needs for analytical tools. People do not always care only about the effect of the explanatory variables on the mean of the response variable, but sometimes they may care about what kind of effect on which level of the response variable. Driven by such practical problems, statisticians have developed quantile regression models, which provide more and more flexible perspectives for people to analyze real-world problems by modeling the conditional quantiles of response variables. Since the quantile regression has been proposed for half a century, its application scenarios have become more and more popular. Compared with mean regression, its superiority has been verified in many practical problems, and it has become a major hotspot of statistical research in recent years.

Since the 21st century, the popularization of computers has brought a great deal of convenience to people's lives and promoted social progress and productivity development, and the explosive growth of computing power in the last decade has pushed the society into the era of big data, where billions of levels of data are generated, transmitted and stored every day, and in the era of big data, it is quite easy to obtain the information related to it, that is to say, the auxiliary information, for the problems of interest. Therefore, how to introduce auxiliary information to improve the efficiency of parameter estimation or prediction accuracy has gradually become one of the hot issues in statistics.

However, in the face of the huge amount of auxiliary information, the randomness of the introduced auxiliary information has to be considered. The introduction of auxiliary information in quantile regression has been explored by researchers from various perspectives, but the introduction of auxiliary information and the consideration of the randomness of the auxiliary information in quantile regression have rarely been studied in this field. Therefore, this paper proposes to introduce auxiliary information into the quantile regression model and consider the stochasticity of auxiliary information, design an algorithm to solve the estimation of quantile regression coefficients, improve the estimation efficiency by introducing auxiliary information, and enhance the robustness of parameter estimation by considering the stochasticity of auxiliary information to bring more information to the model.

In Chapter 3 of this paper, the kernel smoothing method is proposed for the quantile regression model considering the randomness of auxiliary information, which turns the non-smooth quantile loss into a smooth function, thus designing a set of computationally convenient linear iterative algorithms to solve the estimation of regression coefficients. And since the coefficient estimation problem of quantile regression is actually a special kind of parameter estimation problem of U-process, this paper estimates the variance of the regression coefficients using a perturbed resampling method for the objective function, which is a generalized Bootstrap method.

Subsequently, this paper conducts simulation experiments for the proposed algorithm, and the experiments show that the model with stochastic auxiliary information has a smaller bias for parameter estimation compared to the quantile regression model without auxiliary information, with similar variance, and the empirical coverage is closer to the theoretical value, which implies that the algorithm proposed in this paper really improves the efficiency of parameter estimation by introducing the auxiliary information and taking into account its stochasticity and the regression coefficients have a better performance compared to the coefficient estimates without auxiliary information have better asymptotic properties, further illustrating the effectiveness of the algorithm.

The aim of this paper is to provide new perspectives on how to improve the efficiency

of parameter estimation and the problem of introducing auxiliary information by introducing auxiliary information and considering randomness in quantile regression, and to inspire future researchers to explore and innovate more deeply in this field.

**Key Words:** Quantile regression, Auxiliary information, Perturbation resampling, Bootstrap, U-process, Iterative algorithm, Estimation equation

# 第一章　绪论

## 1.1　研究背景与选题意义

统计学是一门能与自然科学和社会科学相结合的综合性学科，主要是通过搜集数据，描述数据和分析数据等手段，实现对数据的理解与探索，挖掘数据背后的规律，因此统计学也被称为数据的科学，更是探索、挖掘和分析数据的艺术。21 世纪以来计算机的广泛应用和算力爆发式的增长，以及数据搜集和储存能力的显著提升，为统计学的快速发展提供了强大的技术支持。尤其是大数据时代的到来，更加使得统计学在处理海量复杂数据集方面发挥日益重要的作用，统计学在各个领域的影响力不断提升，在经济学、生物学、环境科学等重要领域的发展中起到了不可或缺的作用。例如在经济学领域，经济学家使用统计学建立和验证经济模型，对市场趋势进行预测和对经济指标的关系进行评估，为政府制定经济政策提供数据和决策支持；在生物医学领域，统计方法早已被用于临床试验的设计和分析，帮助科学家评估药物的疗效和安全性，推动生命科学和医学研究的进步，近年来由于大数据的兴起，统计学更是被广泛用于分析生物的基因组数据，从基因层面发现生物表型之间的关联，并由此衍生出了生物信息学这一新兴领域。除此之外，在农业领域，农业学家们利用统计学知识研究农作物的生长环境、施肥量多少、土壤状况等因素对农作物产量的影响，从而提高农作物的产量与质量；在保险精算行业中，一些经典统计学模型更是被广泛应用于评估赔付风险，从而确定保单金额和制定保险策略等。

然而，在上述提到的诸多应用场景中，统计学的最初应用大多基于均值回归模型，本质上是对条件期望 $E(Y|X)$ 进行建模，其中 $X$ 表示解释变量，$Y$ 表示响应变量。在现代统计学发展的一百多年来，统计学家们在不同的实际问题的驱动下，在简单均值回归模型的基础上发展衍生出了许多更为复杂的模型，例如，为了捕捉数据中的非线性关

系，发展出了多项式回归模型和非参数回归模型；为了分析时间序列数据，产生了自回归模型 (AR)、移动平均模型 (MA) 以及自回归移动平均 (ARMA) 模型等；而在面对大规模数据时，提出了分布式计算、数据压缩和近似计算等方法以提高计算效率。在最常用的均值回归中，通常选平方损失作为损失函数，在误差满足正态性假定的条件下，求解平方损失得到的最小二乘估计有良好的估计性质，但是最小二乘估计的应用条件比较苛刻，需要误差具有同方差性，甚至要求误差服从正态分布，并且在样本中存在异常值点时，求解平方损失得到的均值回归模型对异常值的表现会非常敏感，当误差是厚尾分布等情况时，最小二乘估计表现欠佳，模型稳健性较差。随着研究的不断深入和应用领域的不断扩大，人们发现现实数据完全满足最小二乘估计所要求的假设的情况并不多，并且一旦数据违背了某一条基本假设，那么通过最小二乘方法就很难得到一个无偏有效的估计量。同时，随着社会的发展，均值回归模型并不能够满足人们的所有需求。人们并不总是只关心解释变量对响应变量均值的影响，在有些时候人们可能更关心解释变量对不同水平的响应变量的影响，例如，在政府制定扶贫政策时，可能会更关心哪些政策对于低收入人群的影响更大；而当政府制定税收政策时，往往可能更关心哪些政策对高收入人群的影响更大。因此，在这样实际问题的驱动下，统计学家提出了分位数回归模型。通过分位数回归，研究者可以选择性地对响应变量的不同分位数水平进行建模。在某种程度上，分位数回归弥补了均值回归的空白，可以更加灵活地挖掘出数据不同层次的关系，并成为近年来统计学研究的一大热点。

分位数回归的概念最早由 Koenker and Bassett(1978)[1]提出，研究解释变量与响应变量条件分位数之间的关系。即对于一个感兴趣的分位数水平 $\tau_0$，我们可以假设如下线性分位数回归模型：

$$Q_{\tau_0}(y|\boldsymbol{x}) = \boldsymbol{x}^{\boldsymbol{T}}\beta_0,$$

其中，$Q_{\tau_0}(y|\boldsymbol{x})$ 表示 $y$ 的条件 $\tau_0$ 分位数，$x$ 表示解释变量，$y$ 表示响应变量。

分位数回归作为一种多角度获取响应变量分布信息的方法，自其诞生起，就在经济学、生物学和环境科学等研究领域广受欢迎。例如，经济金融领域的数据往往表现出一些典型的复杂性特征，例如非平稳性，非线性等，同时这类数据会受到政策、市场环境、技术创新等诸多因素的影响，数据内部常常在不同水平表现出完全不同的特点和规律，这时

采用分位数回归在响应变量的不同分位数水平下“分而治之”，往往更能揭示出数据内部的规律，因此与传统的均值回归方法相比，分位数回归在经济金融领域被更加广泛应用。经济领域的收入不平等问题和金融经济领域的投资组合问题，都有很多学者利用分位数回归方法得到了一些有价值的结论。例如，Buchinsky(1994)[2]将分位数回归应用在收入研究中，研究了具有不同技能群体的收入不平等现象；Baur(2012)[3]利用将分位数回归和自回归模型结合，设定不同的分位数水平，利用分位数自回归方法研究了股票收益的条件分布；而 Wattanawongwan(2023)[4]则利用分位数回归对信用卡数据进行了分析。在经济金融领域分位数回归的应用还有很多，Taylor(1999)[5],Taylor(2000)[6],Hwang(2010)[7]等都是该场景下的实际应用，在这里不过多赘述。

在生物生态学数据中同样存在着与经济金融数据相似的特点，人们发现在自然界中也存在许多非平均的现象，或者存在明显的异方差，例如植物生长数量或者河流污染情况，而分位数回归很适用于处理这类波动性大、异质性强的数据。许多研究人员借助分位数回归模型对生态环境领域数据进行了研究，例如，Koenker and Schorfheide(1994)[8]利用分位数回归分析了全球气候变化的过程，而 Pozarickij et al.(2019)[9]利用分位数回归揭示了近视基因与环境的相互作用关系。

随着大数据时代的不断发展，人们搜集和获取数据的种类和途径越来越多，对于一个感兴趣的问题，往往已有大量的研究从不同角度进行了建模和推断；而对于一个新的独立研究来说，由于各种成本的限制，研究者通常可以获得的内部数据是有限的，并且在一些医学和环境问题当中，获取满足条件的样本本身就是一件耗时耗力且困难的任务，所以如何将研究者的内部数据与已有的外部辅助信息进行融合以提高独立研究的推断效率和有效性，逐渐成为近年来统计学研究的热点问题之一。然而，出于数据隐私和安全性考虑，外部信息往往无法以个体级数据的形式呈现，取而代之的是以摘要信息或者模型的参数估计和统计量的形式呈现。摘要信息比个体级数据更容易获得，一些已发表的文献，研究报告，或者是公共的专业领域数据集往往都会含有摘要信息的报告，这些摘要信息可以作为辅助信息与内部研究相结合，以提高参数的估计效率。然而，随着研究的深入，学者们逐渐意识到，由于摘要辅助信息往往也来源于其他学者或政府机构的研究，所以将辅助信息视作常量是不合适的，尤其是当外部研究的样本量和内部研究的样本量相当时，辅助信息的引入会带来不可忽视的随机性，Zhang et al.(2020)[10]考虑

辅助信息随机性不可忽略的情况下，带有估计方程约束的经验似然的极大似然估计问题。

鉴于此，本文将沿用 Zhang et al.(2020)[10]的思想，考虑在分位数回归框架下，引入带有不可忽略随机性的辅助信息的参数推断问题，并使用核光滑方法将不可导的分位数损失函数转化为可导函数进行求解，并给出相应的迭代算法。本文旨在通过引入辅助变量来增强分位数回归模型的解释力和预测精度，开发新的算法和方法，不仅能够有效地整合辅助信息，并且考虑到了辅助信息不可忽略的随机性，还能够提供关于模型参数的可靠推断。此外，本研究还将关注算法的计算效率和实用性，确保提出的方法可以在实际问题中得到有效应用。本研究的意义在于，不仅推动了分位数回归理论的发展，还为实际问题提供了更为精确和全面的解决方案。通过结合辅助信息，本研究有助于揭示数据中更深层次的内在联系，为决策者提供更为科学的决策依据。此外，本研究还通过数值模拟验证方法的有效性，通过本研究的深入，预期能够促进统计学方法与实际应用的深度融合，为未来的跨学科研究提供新的视角和工具。

## 1.2　文献综述

在这一小节中，将对过去的一些和本文研究领域相关的成果作简要的阐述，旨在通过梳理先前学者的研究成果和方法论，帮助明确本文研究的出发点和方向，并在某种程度上揭示当前领域的发展脉络和未来趋势。

### 1.2.1　分位数回归文献综述

自 Galton(1886)[11]第一次用“回归”(regression) 描述子女身高和父母身高之间的关系起，均值回归在此后近一百年里因其简洁易懂以及可解释性强等优势，在统计分析中占据了重要的席位。然而随着社会进步和研究的不断深入，统计方法应用的场景和人们关心的问题也变得越来越复杂，尤其当数据内部呈现出分层结构时，均值回归往往会忽视这种分层结构。并且均值回归对模型误差的假设过于严格，尽管后续的学者尝试放宽独立同分布的假设，但在实际当中对误差项施加的参数化假设仍然难以得到满足。例如，White(1980)[12]提出了一种异方差情况下回归系数的协方差矩阵的估计方法，并指出了 OLS 在面对异方差问题时的局限性；Huber(1987)[13]讨论了在非标准最小二乘条件

下，回归系数的极大似然估计的行为，为后续放宽 OLS 的严格假设提供了理论基础；而 Carroll(2017)[14]讨论了回归分析中的变换和加权方法，这些方法可以用来放宽 OLS 的假设。

在实际问题和均值回归局限性的驱动下，在最小一乘 (LAD) 的基础上，Koenker and Bassett(1978)[1]首次理论化地提出了分位数回归的想法，最小一乘 (LAD) 是分位数回归的一种特殊情况，当分位数水平为 0.5 时，分位数回归问题就退化到 LAD 问题。因此，Koenker and Bassett(1978)[1]的工作可以看做是对 LAD 的一种推广，将 0.5 分位数的特殊情况推广至任意分位数水平。分位数回归放宽了对模型误差的限制，不对模型误差做任何参数化的假设，仅要求误差的 $\tau$ 分位数为 0，这一要求可以通过调整模型的截距项自然满足 (Koenker, 1978)[1]。分位数回归相对于 OLS 回归有很多优势，因其是对条件分位数而不是条件均值建模，所以在处理厚尾数据方面有着天然的优势，相比基于条件均值的方法对异常值更加稳健，更能详细地描述变量的统计分布。

在分位数回归发展的历程中，Koenker and Bassett(1982)[15]研究了分位数回归对异方差的稳定性和分位数回归的线性假设检验问题，同时 Bassett and Koenker(1986)[16]证明了分位数回归估计的强相合性，为分位数回归奠定了理论基础。在此之后，Buchinsky(1995)[17]研究了分位数回归系数的渐进协方差矩阵的估计。

### 1.2.2　辅助信息文献综述

分位数回归在一定程度上填补了均值回归的空白，它允许研究者在不同的分位数水平上对响应变量进行建模，能够更好地处理数据内部的复杂结构和异质性。然而，在实际的数据分析问题当中，我们常常会遇到包含丰富的辅助信息 (记作 $\tilde{\beta}$) 的数据集，或者是所关心的问题已有很多相关的研究，而从这些相关的研究结果中可以提取到丰富的与当前的问题相关的辅助信息用以辅助当前问题的推断。在实际问题中，研究者通常基于当前研究的内部数据进行统计推断，而希望借用基于外部数据得到的辅助信息提高当前研究的推断效率。通常我们只能获得经过总结后得到的摘要信息，而无法获得外部研究的个体级数据 (Zhang et al. 2020)[18]。Lin and Zeng(2010)[19]指出当内部研究和外部研究采用同样的模型时，传统的 meta-analysis 可以将各研究的数据混合以提高推断效率。然而在实际当中，研究者通常采用不同的模型研究相似的问题 (Zhang, 2020)[18]，近年来，

许多学者分别从频率学派和贝叶斯学派的观点出发，尝试改进传统的 meta-analysis，希望充分利用基于不同协变量甚至是不同似然函数得到的信息。

其中，Chen and Qin(1993)[20]研究了有限总体下结合辅助信息的经验似然推断问题；Chen and Sitter(1999)[21]研究了经验似然结合复杂调查中的辅助信息的推断问题，Imbens and Lancaster(1994)[22]在微观计量领域，通过融合微观与宏观数据，提高微观计量模型的解释力和预测准确性；Chen et al.(2002)[23]使用经验似然给出的权重分布作为辅助信息，提高加权回归估计器的表现；而 Cheng et al.(2018a)[24]从贝叶斯学派的观点出发，研究了如何将外部系数作为辅助信息整合到二元结果的风险预测模型中；Cheng et al.(2018b)[25]同样从贝叶斯的角度出发，研究了在线性回归中如何整合来自一个已经建立的简化模型的外部信息。在引入 $\tilde{\beta}$ 时，若外部研究的样本量比内部研究的样本量大得多，则 $\tilde{\beta}$ 的随机性可以忽略 (Zhang et al. 2020)[18]。Qin(2000)[26]提出了一种在经验似然中将内外部参数的估计方程作为辅助信息整合的方法，但当 $\tilde{\beta}$ 的随机性不可忽略时，Zhang et al.(2020)[18]提出了一种在参数模型下，考虑 $\tilde{\beta}$ 随机性的整合辅助信息的方法。

### 1.2.3　Bootstrap 文献综述

Bootstrap 是一种统计重抽样方法，由 Efron(1979)[27]提出，该方法不依赖于对总体的任何假定，而是通过对已有样本进行有放回抽样获得新的 Bootstrap 样本，基于 Bootstrap 样本对感兴趣的参数进行统计推断。随后 Efron(1981a)[28]，Efron(1981b)[29]，Efron(1982)[30]进一步拓展了 Bootstarp 方法的适用范围和思想。因为 Bootstrap 不需要对观测数据假定具体的分布形式，并且在有限样本的情况下表现也比较好，所以 Bootstrap 方法自提出以来，因其广泛的适用性，在理论和应用研究方面得到了充分的发展。迄今为止已经发展出了多种 Bootstrap 方法，其中，常见的 Bootstrap 方法包括：残差 Bootstrap、参数 Bootstrap、Wild Bootstrap、Pairs Bootstrap 和 Block Bootstrap 等等。

残差 Bootstrap 即对模型残差进行 Bootstrap 抽样，要求模型误差与解释变量相互独立，并且模型误差独立同分布 (欧变玲, 2011)[31]，残差 Bootstrap 方法是一种最基本的非参数 Bootstrap 方法 (Efron, 1979)[27]，可以用于横截面数据和面板数据的分析，但不能处理时间序列数据；参数 Bootstrap 利用先验信息确定样本数据总体的分布，然后从总体分布中抽取额外的样本，这是一种参数方法，其思想与蒙特卡罗方法类似，二者都是

从已知的数据生成过程中进行抽样；Wu(1986)[32]提出了 Wild Bootstrap 方法，适用于处理异方差问题，Liu(1988)[33]和 Mammen(1993)[34]对其进行了拓展，分别讨论了非独立同分布误差下的 Bootstrap 抽样问题和高维线性模型的 Bootstrap 抽样问题；Pairs Bootstrap 方法不是对模型残差抽样，而是成对地抽取样本观测值，由 Freedman(1981)[35]和 Freedman(1984)[36]提出，后由 Flachaire(1999)[37]改进将其应用于异方差的非线性模型；Block Bootstrap 是一种处理相关性数据的方法，其基本思想是分块对样本进行 Bootstrap 抽样，Politis(2003)[38]将其应用于时间序列数据，Gonçalves et al.(2004)[39]和 Horowitz(2003)[40]对 Block Bootstrap 的有效性条件进行了进一步研究。

Hu and Kalbfleisch(2000)[41]提出了一种直接扰动得分函数产生 Bootstrap 样本进行推断的方法，然而，在非光滑损失的情况下使用 Hu and Kalfleisch(2000)[41]的方法进行扰动重抽样并计算参数估计值是较为困难的 (Jin et al. 2001)[42]。为了解决非光滑损失的情况，Jin et al.(2001)[42]提出了一种新的采样方法，适用于估计 U-统计量核函数 h 的参数 $\theta$，该方法通过直接扰动目标函数而非得分函数来估计参数的分布，适用于非光滑目标函数和数据生成过程未知的情况，并且和 Bootstrap 一样具有广泛适用且步骤简单、清晰易懂的特点。本文将这种重采样方法平移到具有非光滑损失的分位数回归当中，通过扰动目标函数来估计分位数回归系数的方差。

## 1.3　论文结构

在本文的第二章，为了后续研究工作的展开，我们将本文研究的模型及基础研究方法做简要介绍，主要包括了分位数回归的基本理论，核函数近似以及 Bootstrap 的主要思想等，这些内容会为后续的研究提供背景支持。

在第三章中，我们将给出问题的定义，求解目标函数并给出迭代算法。

在第四章中，我们将进行数值模拟，汇报算法模拟的结果并与其他方法进行比较。

在文章最后，将对全文进行总结，列举出研究进行的创新点与不足，讨论并提出未来可进一步研究的方向。

# 第二章　模型介绍

## 2.1　分位数回归模型

分位数回归模型由 Koenker and Bassett(1978)[1]提出，相较于均值回归，分位数回归在处理异常值时更加稳健，并且设定不同的分位数水平，可以得到响应变量的条件分布的更加全面的信息。分位数回归自提出以来，广泛应用于经济、金融、管理、生物医学和环境科学等领域，并进一步发展出了更加复杂的分位数回归模型。本文考虑最一般的线性分位数回归模型，在这里简要介绍分位数回归模型的基本思想。

考虑实值随机变量 $X$，具有分布函数 $F(x) = P(X \leq x)$，对于任意给定的实数 $\tau : 0 < \tau < 1$，随机变量 $X$ 的 $\tau$ 分位数定义为:

$$Q(\tau) = \inf\{x : F(x) > \tau\}, \tag{2.1}$$

考虑响应变量 $Y$，解释变量 $X$，设 $F_{Y|X} = P(Y \leq y|X)$ 表示给定 $X$ 条件下 $Y$ 的条件分布，给定分位数水平 $\tau \in (0,1)$，则分位数回归实际上是在对条件分布 $Y|X$ 的 $\tau$ 分位数进行建模，其中响应变量 $Y$ 的条件 $\tau$ 分位数由下式定义:

$$Q_{Y|X}(\tau) = \inf\{y : F_{Y|X}(y) > \tau\}.$$

分位数回归的损失函数定义为:

$$\rho_\tau(u) = u\{\tau - I(u < 0)\}, \tag{2.2}$$

这是因为对于任意随机变量 $X$，其 $\tau$ 分位数 $(0<\tau<1)$ 具有如下性质：

$$Q_X(\tau)=\arg\min_c \mathrm{E}\left\{\rho_\tau(X-c)\right\}, \tag{2.3}$$

仅对 $X$ 为连续型随机变量的情况对上述性质做简要说明，设 $X$ 为连续型随机变量，分布函数为 $F(x)$，密度函数为 $f(x)$，则有：

$$\begin{aligned}
0&=\frac{\partial}{\partial c}\mathrm{E}\rho_\tau(X-c)=\frac{\partial}{\partial c}\int_R\rho_\tau(x-c)f(x)\,dx\\
&=\frac{\partial}{\partial c}\int_R(x-c)(\tau-I(x-c<0))f(x)\,dx\\
&=\frac{\partial}{\partial c}\left[\int_{-\infty}^{c}(x-c)(\tau-1)f(x)\,dx+\int_c^{+\infty}(x-c)\tau f(x)\,dx\right]\\
&=(1-\tau)\int_{-\infty}^{c}f(x)\,dx-\tau\int_c^{+\infty}f(x)\,dx\\
&=(1-\tau)F(c)-\tau(1-F(c))\\
&=F(c)-\tau.
\end{aligned}$$

所以

$$\begin{aligned}
&\tau=F(c),\\
&c=F^{-1}(\tau)\overset{\text{def}}{=}\inf\{x:F(x)>\tau\}.
\end{aligned}$$

注意到损失函数 $\rho_\tau$ 还可以等价地写成下面的形式：

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

由此可以看出分位数回归的损失函数 $\rho_\tau$ 是一个对钩形凸函数，且仅在 $u=0$ 点不可导。

在分位数回归当中，我们是想用解释变量 $\boldsymbol{X}$ 的一个线性函数 $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta}$ 来拟合响应变量 $Y$ 的 $\tau$ 分位数，根据性质 (2.3)，我们知道满足条件的 $\boldsymbol{\beta}$ 应当由下式给出：

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \mathrm{E}\left\{\rho\left(Y - \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta}\right)\right\},$$

其样本形式为：

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n} \rho\left(y_i - \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{\beta}\right), \tag{2.4}$$

(2.4) 就是分位数回归中回归系数解的一般形式。

同时应当注意到分位数回归和最小一乘回归 (LAD) 之间的关系：当 $\tau = 0.5$ 时，分位数回归的损失函数就退化绝对离差损失，此时分位数回归问题就退化到最小一乘回归问题。从这个角度来看，一般的分位数回归也可以看做是 LAD 的推广，更具体地，最小一乘回归 (LAD) 方法是使得误差的绝对值之和最小，即

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} |y_i - \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{\beta}|, \tag{2.5}$$

而分位数回归则是在 LAD 的基础上，赋予不同等级的绝对离差以一定的权重，最小化他们的加权绝对离差和，也就是说分位数回归实际上是 LAD 的一种加权版本。给定分位数水平 $\tau$，我们有 (Koenker and Bassett, 1978)[1]：

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \left( (1-\tau) \sum_{i:y_i < \boldsymbol{x_i}^T\boldsymbol{\beta}} |y_i - \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{\beta}| + \tau \sum_{i:y_i \geq \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{\beta}} |y_i - \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{\beta}| \right) \\ &= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\tau\left(y_i - \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{\beta}\right). \end{aligned} \tag{2.6}$$

本文考虑最一般的线性分位数回归，即假设响应变量 $Y$ 的条件 $\tau$ 分位数 $Q_{Y|\boldsymbol{X}}(\tau)$ 与解释变量 $\boldsymbol{X}$ 之间是线性关系。下面给出线性分位数回归的一般形式，设解释变量 $\boldsymbol{X} = (1, X_1, X_2, ..., X_p)^T \in \mathbf{R}^{p+1}$，响应变量 $Y \in R^1$，回归系数 $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p) \in \mathbf{R}^{p+1}$，记 $Q_{Y|\boldsymbol{X}}(\tau)$ 表示给定 $\boldsymbol{X}$ 条件下 $Y$ 的条件 $\tau$ 分位数，则线性分位数回归假设如下模型:

$$Y = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta} + \epsilon;\ Q_{\epsilon|\boldsymbol{X}}(\tau) = 0. \tag{2.7}$$

或者等价地：

$$Q_{Y|\boldsymbol{X}}(\tau) = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta}(\tau), \tag{2.8}$$

其中 $Q_{\epsilon|\boldsymbol{X}}(\tau)$ 表示给定 $\boldsymbol{X}$ 情况下 $\epsilon$ 的条件 $\tau$ 分位数。回归系数的理论解 $\boldsymbol{\beta}_h$ 为：

$$\boldsymbol{\beta}_h = \arg\min_{\boldsymbol{\beta}} \mathrm{E}\left\{\rho_\tau\left(Y - \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta}\right)\right\}, \tag{2.9}$$

在样本的经验分布 $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X \le x\}$ 下，得到其样本类似：

$$\widehat{\boldsymbol{\beta}}_h = \arg\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\left(Y_i - \boldsymbol{X}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{\beta}\right). \tag{2.10}$$

分位数回归相较于均值回归，具有许多优良的性质，在这里我们给出最重要的几点：

1. 在分位数回归中，对于模型的随机误差项不做任何参数性假设，仅要求 $Q_{\epsilon|\boldsymbol{X}}(\tau) = 0$，通过改变截距项的真值可以使该条件恒成立。
2. 分位数回归系数对异常数据点不敏感，表现出很强的稳健性。
3. 分位数回归对于响应变量具有单调不变性，即对响应变量做单调变换不改变分位数回归的系数。
4. 分位数回归可以选择多个分位数水平进行研究，因此分位数回归比均值回归更能得到数据更深层更全面的信息。

分位数回归已经被证明具有良好的渐近性质，对于一个单变量的分位数回归模型（即只考虑截距项），我们给出其回归系数的如下性质 (Koenker and Bassett, 1978)[1]：

**定理 2.1.** 设 $\{y_1, y_2, ..., y_n\}$ 是独立同分布的样本，分布函数为 $F$，分布函数一阶可导，密度函数为 $f$，则其 $\tau$ 分位数由下式给出：

$$\widehat{\alpha}_\tau = \arg\min_{\alpha} \frac{1}{n}\sum_{i=1}^{n} \rho\left(y_i - \alpha\right),$$

且满足：

$$\lim_{n\to\infty} \sqrt{n}(\widehat{\alpha}_\tau - \alpha_\tau) \xrightarrow{d} N\left(0, \frac{\tau(1-\tau)}{f^2(F^{-1}(\tau))}\right),$$

其中 $d$ 表示依分布收敛。

## 2.2 Bootstrap 重抽样

Bootstrap 方法又称自助法，是一种统计重抽样技术，由 Efron(1979)[27]提出，它不需要对未知总体做任何形式的假定，而是通过对已有样本进行有放回的随机抽样得到一组 Bootstrap 样本，基于 Bootstrap 样本对关心的未知参数进行统计推断。使用 Bootstrap 方法进行统计推断的一般步骤如下：

1. 已知存在一组样本 $\boldsymbol{X} = (X_1, X_2, ..., X_n)^T$ 来自某一未知分布 $F$，现有一组样本的观测值 $\boldsymbol{x} = (x_1, x_2, ..., x_n)^T$，假设我们统计推断关心的参数是 $\theta$，是某个依赖于样本 $X_i$ $(i = 1, ..., n)$ 和总体分布 $F$ 的统计量。

2. 使用特定的 Bootstrap 方法抽取 Bootstrap 样本 $B$ 次，得到 $B$ 个 Bootstrap 样本，记作 $X^{*1}, X^{*2}, ..., X^{*B}$.

3. 基于 $B$ 个 Bootstrap 样本，可以构造 $B$ 个经验分布 $F^{*1}, F^{*2}, ..., F^{*B}$.

4. 基于经验分布 $F^{*i}$ 求出 $B$ 组 Bootstrap 样本 $\theta$ 的估计值，记作 $\theta_1^*, \theta_2^*, ..., \theta_B^*$.

5. 则 $\theta$ 的估计值为：
$$\widehat{\theta} = \frac{1}{B}\sum_{i=1}^{B}\theta_i^*. \tag{2.11}$$

6. 方差 $\sigma_\theta$ 的估计值为：
$$\widehat{\sigma}_\theta^2 = \frac{1}{B-1}\sum_{i=1}^{B}\left(\theta_i^* - \widehat{\theta}\right)^2. \tag{2.12}$$

## 2.3 U-统计量

U-统计量最早由 Hoeffding(1948)[43]提出，是一类特定的，具有对称性质的统计量，U-统计量都是无偏的，并且通常具有良好的渐进正态性。在非参数统计和半参数统计中，许多用于估计和检验的统计量在实质上都是 U-统计量，U-统计量近年来在了解复杂随机过程和随机网络类型数据的性质方面发挥了重要作用。

下面我们先给出一般的 U-统计量的一般定义，然后说明样本分位数可以在特定情况下由 U-统计量导出，可以看做是 U-统计量的一个特殊情况。

**定义 2.2.** 设 $h(x_1, x_2, ..., x_n)$ 为任意多元实值函数，$h : \mathbf{R}^n \to \mathbf{R}$，若 $h$ 满足：对于任意 $\{1, 2, ..., n\}$ 的置换 $\sigma$（即 $\sigma$ 是 $\{1, 2, ..., n\}$ 的任意排列），都有 $h(x_1, x_2, ..., x_n) = h(x_{\sigma(1)}, x_{\sigma(2)}, ..., x_{\sigma(n)})$，则称 $h$ 为对称函数 (symmetric function).

**定义 2.3.** 设 $h(x_1, x_2, ..., x_m)$ 为任意实值可测的对称函数，设 $X_1, X_2, ..., X_n$ 为来自概率分布 $F$ 的样本且 $n \geq m$，则核为 $h$ 的 $m$ 阶 U 统计量定义为：

$$U_n = U_n(h) = \frac{1}{\binom{n}{m}} \sum_{\mathbf{C}_{m,n}} h(X_{i_1}, \ldots, X_{i_m})$$

其中 $\mathbf{C}_{m,n}$ 表示所有可能的来自 $\{1, 2, ..., n\}$ 的大小为 $m$ 的组合

接下来介绍样本分位数如何是 U-统计量的特殊情况，假设总体分布为 $F$, 在抽取一组样本后，由样本构造的经验分布为 $F_n$，对于任意实数 $t$，注意到当 $m = 1$, $h(x) = I\{x \leq t\}$，总体参数 $\theta = P(X \leq t)$ 时，U 统计量退化为如下形式，它实际上就是经验分布 $F_n$ 在点 $t$ 处的取值：

$$U_n^*(t) = F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I\{x \leq t\}. \tag{2.13}$$

对于给定的分位数水平 $\tau$，由 (2.1) 及 (2.13) 知，样本 $\tau$ 分位数为：

$$\widehat{Q}(\tau) = \inf\{t : U_n^*(t) > \tau\} \tag{2.14}$$

并且由性质 (2.4)，我们知道由 (2.13) 所定义的样本分位数一定满足下式：

$$\widehat{Q}(\tau) = \arg\min_c \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(X_i - c). \tag{2.15}$$

在分位数回归问题当中，我们实际上是用给定 $\boldsymbol{X} = \boldsymbol{x}$ 的条件下，$Y$ 的样本条件分位数 $\boldsymbol{x}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}$ 作为统计量去估计给定 $\boldsymbol{X} = \boldsymbol{x}$ 的条件下，$Y$ 的总体条件分位数 $Q_{Y|\boldsymbol{X}=\boldsymbol{x}}(\tau)$。从这个角度来看，分位数回归系数的估计问题就转变成一个核函数带未知参数的 1 阶 U 统计量的参数估计问题。

取 $m=1$, $h(y)=I\left\{y \leq \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{\beta}(\tau) \mid \boldsymbol{X}=\boldsymbol{x}\right\}$, $\theta=P\left\{Y \leq \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{\beta}(\tau) \mid \boldsymbol{X}=\boldsymbol{x}\right\}$, 则 U-统计量变为:

$$U_n\Big(\boldsymbol{\beta}(\tau)\Big)=\frac{1}{n} \sum_{i=1}^{n} I\left\{y \leq \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{\beta}(\tau) \mid \boldsymbol{X}=\boldsymbol{x}\right\}, \tag{2.16}$$

它是在模型假设 (2.7) 下，在给定 $\boldsymbol{X}=\boldsymbol{x}$ 条件下，$Y$ 的经验分布，并由此得到 $Y$ 的条件 $\tau$ 分位数的样本估计:

$$\widehat{Q}_{Y \mid \boldsymbol{X}=\boldsymbol{x}}(\tau)=\inf\left\{\boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{\beta}(\tau): U_n\Big(\boldsymbol{\beta}(\tau)\Big)>\tau\right\}, \tag{2.17}$$

以及在模型假设 (2.7) 下，回归系数的估计值 $\widehat{\boldsymbol{\beta}}(\tau)$:

$$\widehat{\boldsymbol{\beta}}(\tau)=\inf\left\{\boldsymbol{\beta}(\tau): U_n\Big(\boldsymbol{\beta}(\tau)\Big)>\tau\right\} \tag{2.18}$$

并注意到式 (2.18) 一定满足:

$$\widehat{\boldsymbol{\beta}}(\tau)=\arg\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n}\left\{\rho_\tau\left(Y-\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{\beta}\right)\right\}. \tag{2.19}$$

# 第三章　带随机性辅助信息的分位数回归

## 3.1　模型介绍

在这一小节，我们首先介绍本文研究的带辅助信息的分位数回归模型的具体设定，其次介绍估计回归系数的方差时用到的扰动重抽样方法 (Jin et al. 2001)[42]。

### 3.1.1　分位数回归模型介绍

假设有一组来自如下模型的独立同分布 (Independent Identically Distributed，i.i.d.) 样本 $\{(\boldsymbol{X_i}, Y_i)\}$:

$$Y_i = \boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}(\tau) + \epsilon_i,\ i = 1, ..., n \tag{3.1}$$

其中 $\boldsymbol{X_i}^{\boldsymbol{T}} = (1, X_{i1}, ..., X_{ip}) \in \mathbf{R}^{p+1}$ 是从总体 $X$ 中抽样得到的简单随机样本，误差项 $\epsilon_i$ 是一个无法观测的随机变量，对于给定的分位数水平 $0 < \tau < 1$，$\epsilon_i$ 满足 $P(\epsilon_i \leq 0|\boldsymbol{X_i}) = \tau$，即 $\boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}(\tau)$ 是给定 $\boldsymbol{X_i}$ 条件下 $Y$ 的条件 $\tau$ 分位数。

可以通过求解以下极小化问题得到分位数回归系数的估计：

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbf{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}), \tag{3.2}$$

但是注意到分位数损失函数 $\rho_\tau(u) = u(\tau - I\{u < 0\})$ 是一个非光滑的函数，所以直接求解上述问题是不方便的，本文采用核光滑的方法，用一个光滑函数逼近示性函数 $I(\cdot)$，将 (3.2) 转化为一个光滑函数，从而可以通过求导和极值一阶条件得到问题的解。

### 3.1.2　辅助信息模型介绍

在本文中，考虑来自外部的辅助信息 $\widetilde{\boldsymbol{\beta}}$ 以摘要信息的方式呈现，即我们只知道一些摘要统计量的值，而不知道这些摘要统计量是如何得到的。$\widetilde{\boldsymbol{\beta}}$ 是针对同样的总体 $(X, Y)$，基于外部研究的数据得到的分位数回归系数，本文假设外部样本量 $N$ 已知，考虑外部样本量 $N$ 和内部样本量 $n$ 相当的情况，此时辅助信息 $\widetilde{\boldsymbol{\beta}}$ 的随机性不可忽略 (Zhang et al. 2020)[10]，我们用 $\boldsymbol{\Sigma}$ 表示辅助信息 $\widetilde{\boldsymbol{\beta}}$ 的协方差矩阵，在实际当中总体的协方差矩阵 $\boldsymbol{\Sigma}$ 是未知的，因此我们需要用矩阵 $\widehat{\boldsymbol{\Sigma}}$ 对其进行估计，并且 $\widehat{\boldsymbol{\Sigma}}$ 对 $\boldsymbol{\Sigma}$ 估计得越好，引入辅助信息 $\widetilde{\boldsymbol{\beta}}$ 对参数估计效率的提升就越明显 (Zhang et al. 2020)[10]。在本文中，简单地选取 $\widehat{\boldsymbol{\Sigma}} = \mathbf{I}_{p+1} \in \mathbf{R}^{(p+1)\times(p+1)}$，即用 $p+1$ 维的单位阵作为辅助信息协方差矩阵的估计。

假设外部研究的个体级数据为 $\{(\boldsymbol{\phi}(\boldsymbol{X_i}), Y_i),\ i = n+1, ..., n+N\}$，其中 $\boldsymbol{\phi}(\boldsymbol{X})$ 是一已知函数，假设外部研究所使用的估计方法已知，由函数 $g(\cdot)$ 所定义，考虑辅助信息 $\widetilde{\boldsymbol{\beta}}$ 是如下估计方程的解的情况：

$$\sum_{i=n+1}^{n+N} g\{\boldsymbol{\phi}(\boldsymbol{X_i}), Y_i; \boldsymbol{\alpha}, \boldsymbol{\beta}\} = 0, \tag{3.3}$$

其中 $\alpha$ 是冗余参数，即在外部研究中得到的，但与内部研究无关或者无法获得的参数。

在使用后文提到的核光滑方法将分位数损失函数用一个光滑可导函数 $K_h(\cdot)$ 近似后，本文假设函数 $g\{\boldsymbol{\phi}(\boldsymbol{X_i}), Y_i; \boldsymbol{\alpha}, \boldsymbol{\beta}\} = \left\{\partial K_h\left(Y_i - \boldsymbol{X^T\beta}\right)\right\}/\partial\boldsymbol{\beta}$，则由 (3.3) 解得的 $\boldsymbol{\beta}$ 就是在同样的总体 $(\boldsymbol{X}, Y)$ 下，外部辅助信息对于分位数回归系数的估计。

从假设数据的真实分布为 $f(\boldsymbol{X}, Y; \boldsymbol{\theta}_0) = f(\boldsymbol{X}) f(Y \mid \boldsymbol{X}; \boldsymbol{\theta}_0)$，其中 $\boldsymbol{\theta}_0$ 是内部研究关心的总体参数 $\boldsymbol{\theta}$ 的真值，令 $\boldsymbol{\mu} = (\boldsymbol{\theta^T}, \boldsymbol{\alpha^T}, \boldsymbol{\beta^T})^T$，设 $\boldsymbol{\mu_0} = (\boldsymbol{\theta_0^T}, \boldsymbol{\alpha_0^T}, \boldsymbol{\beta_0^T})^T$ 为所有参数的真值，假设 $\widetilde{\boldsymbol{\alpha}}$, $\widetilde{\boldsymbol{\beta}}$ 是 (3.3) 的解，White(1982)[44]证明了 $\widetilde{\boldsymbol{\alpha}}$ 和 $\widetilde{\boldsymbol{\beta}}$ 的正态性：

$$N^{1/2}\begin{pmatrix} \widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 \\ \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix} \to N\left(0, A^{-1}BA^{-1}\right).$$

其中

$$A=\int_{\boldsymbol{X}}\left[\int_{Y}\frac{\partial h\{\boldsymbol{\phi}(\boldsymbol{X}),Y;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0\}}{\partial(\boldsymbol{\alpha},\boldsymbol{\beta})}f(Y\mid\boldsymbol{X};\boldsymbol{\theta}_0)\mathrm{d}Y\right]f(\boldsymbol{X})\mathrm{d}\boldsymbol{X},$$
$$B=\int_{\boldsymbol{X}}\left[\int_{Y}h\{\boldsymbol{\phi}\left(\boldsymbol{X}\right),y;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0\}h^{\mathrm{T}}\{\boldsymbol{\phi}\left(\boldsymbol{X}\right),Y;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0\}f\left(Y\mid\boldsymbol{X};\boldsymbol{\theta}_0\right)\mathrm{d}Y\right]f\left(\boldsymbol{X}\right)\mathrm{d}\boldsymbol{X}.$$

若设 $A^{-1}BA^{-1}$ 中对应于 $\widetilde{\boldsymbol{\beta}}$ 的子矩阵为 $\boldsymbol{\Sigma}$，则 $N^{-1}\boldsymbol{\Sigma}$ 即为 $\widetilde{\boldsymbol{\beta}}$ 的协方差矩阵。

基于 $\widetilde{\boldsymbol{\beta}}$ 的渐进正态性，本文考虑用其对数似然函数 $-N(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\boldsymbol{\Sigma}_0^{-1}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta})/2$ 来刻画 $\widetilde{\boldsymbol{\beta}}$ 的随机性，并将其整合到分位数回归的损失函数中，得到本文优化的目标函数：

$$\mathcal{L}\left(\boldsymbol{\beta}\right)=\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(Y_i-\boldsymbol{X_i^T}\boldsymbol{\beta}\right)-N\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^T\boldsymbol{\Sigma}^{-1}\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)/2. \tag{3.4}$$

### 3.1.3　基于 U-过程的扰动重抽样介绍

在 (2.3) 中，已经介绍了 U-统计量的一般定义。特别的，当 U-统计量的核函数带有未知参数 $\theta$ 时，由于 $\theta$ 具有随机性，所以此时的 U-统计量可以看做是一个随机过程，叫做 U-过程。一般来说，设 $\{Z_i\}_{i=1}^n$ 是一列独立同分布的随机变量，$\theta$ 是参数空间 $\vartheta\subset\mathbf{R}^r$ 中的值，则一个核为 $h$ 的 $K$ 阶 U-过程的定义如下：

**定义 3.1.**

$$U_n(\theta)=\binom{n}{K}^{-1}\sum_{1\leqslant i_1<i_2<\ldots<i_K\leqslant n}h(Z_{i_1},\ldots,Z_{i_K};\theta), \tag{3.5}$$

其中 $h(\cdot)$ 是对称函数，$(i_1,...,i_K)\subset\{1,...,n\}$

由 (3.5) 可知，给定 $\theta$，U-过程就是一个核函数为 $h$ 的 $K$ 阶 U-统计量。设 $\widehat{\theta}$ 是 (3.5) 的极小值点，当核函数 $h\left(\cdot\right)$ 是非光滑函数时，求解 $\widehat{\theta}$ 的协方差矩阵比较困难。Jin et al.(2001)[42]提出通过对 U-过程目标函数加扰动的方式，求解扰动后的目标函数，得到扰动重抽样的估计量，用扰动重抽样得到的估计的协方差矩阵近似 $\widehat{\theta}$ 的协方差矩阵。

更具体的，设 $\{z_i\}_{i=1}^n$ 是变量 $\{Z_i\}_{i=1}^n$ 的一组观测值，设 $\{V_i\}_{i=1}^n$ 独立同分布于一个非负且分布已知的随机变量 $V$，均值为 $\mu$，方差为 $K^2\mu^2$. 考虑 $U_n\left(\theta\right)$ 的一个随机扰动

$\widetilde{U}_n(\theta)$：

$$\widetilde{U}_n(\theta) = \binom{n}{K}^{-1} \sum_{1\leqslant i_1<i_2<\ldots<i_K\leqslant n} (V_{i_1}+\ldots+V_{i_K})h(z_{i_1},\ldots,z_{i_K};\theta). \tag{3.6}$$

注意到对于给定的真值 $\theta_0$，(3.6) 的随机性仅来自于 $\{V_i\}_{i=1}^n$，设 $\theta^*$ 是 (3.6) 的极小值点，Jin et al.(2001)[42]证明了在特定的条件下，$n^{\frac{1}{2}}(\theta^*-\widetilde{\theta})$ 是 $n^{\frac{1}{2}}(\widehat{\theta}-\theta_0)$ 的一个好的近似，其中 $\widetilde{\theta}$ 是 $\widehat{\theta}$ 的样本观测值，即在样本 $\{z_i\}_{i=1}^n$ 下，通过最小化 (3.5) 得到的 $\theta$ 的估计值。

通过这种扰动重抽样的方法估计 $\widehat{\theta}$ 的协方差矩阵，其步骤与 Bootstrap 方法非常一致，一般来说分为以下几步：

1. 产生 $B$ 个 $\{V_i\}_{i=1}^n$ 的随机样本，对于第 $j$ 个样本，最小化 (3.6) 获得估计值 $\theta_j^*$.
2. 计算 $\{\theta_j^*\}_{j=1}^B$ 的样本协方差矩阵，将其作为 $\widehat{\theta}$ 协方差矩阵的估计。

## 3.2　公式推导

### 3.2.1　回归系数估计公式

本文将分位数回归损失函数与辅助信息的似然函数结合起来得到目标函数 (3.4)：

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(Y_i-\boldsymbol{X_i^T}\boldsymbol{\beta}\right) - N\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^T\boldsymbol{\Sigma}^{-1}\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)/2.$$

目标是设计算法求解极值问题，得到回归系数 $\widehat{\boldsymbol{\beta}}$ 的估计：

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbf{R}^{p+1}}{\arg\min}\,\mathcal{L}(\boldsymbol{\beta}). \tag{3.7}$$

注意到分位数回归的损失函数是非光滑函数，主要是因为示性函数 $I(\cdot)$ 是非光滑的，使用一光滑函数 $H\left(\frac{x}{h}\right)$ 近似示性函数 $I\{x\geq 0\}$，其中 $h$ 是窗宽，$H\left(\frac{x}{h}\right)$ 应满足如下性质 (Chen et al. 2019)[45]：

1. 当 $u\geq 1$ 时，$H(u)=1$；当 $u\leq -1$ 时，$H(u)=0$,
2. 光滑函数 $H(\cdot)$ 二阶可导，二阶导数 $H^{(2)}(\cdot)$ 有界，且窗宽 $h=o(1)$.

在本文中，$H(\cdot)$ 选做如下函数，可以轻易验证其满足上述条件：

$$H(v)=\begin{cases}0 & \text{if } v\leq -1,\\ \frac{1}{2}+\frac{15}{16}\left(v-\frac{2}{3}v^3+\frac{1}{5}v^5\right) & \text{if } |v|<1,\\ 1 & \text{if } v\geq 1.\end{cases} \tag{3.8}$$

用光滑函数 $H(\cdot)$ 近似示性函数 $I(\cdot)$ 后，可以得到分位数损失的一个光滑近似：

$$K_h(x)=x\left\{H\left(\frac{x}{h}\right)+\tau-1\right\}.$$

则目标函数变为：

$$L(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^{n}K_h\left(Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}\right)-N\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^T\boldsymbol{\Sigma}^{-1}\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)/2. \tag{3.9}$$

优化目标变为：

$$\widehat{\boldsymbol{\beta}}_h=\underset{\boldsymbol{\beta}\in\mathbf{R}^{p+1}}{\arg\min}\,L(\boldsymbol{\beta}). \tag{3.10}$$

由极值的一阶条件，(3.10) 一定满足 $\frac{\partial L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=0$，从而有：

$$\sum_{i=1}^{n}\boldsymbol{X_i}\left\{H\left(\frac{Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}}{h}\right)+\tau-1+\frac{Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}}{h}H'\left(\frac{Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}}{h}\right)\right\}+nN\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)=0. \tag{3.11}$$

记

$$\begin{aligned}\boldsymbol{U}=&\sum_{i=1}^{n}\boldsymbol{X_i}\left\{H\left(\frac{Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_0}{h}\right)+\tau-1+\frac{Y_i}{h}H'\left(\frac{Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_0}{h}\right)\right\}\\&-\sum_{i=1}^{n}\boldsymbol{X_i}\boldsymbol{X_i}^{\boldsymbol{T}}\frac{1}{h}H'\Big(\frac{Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_0}{h}\Big)\widetilde{\boldsymbol{\beta}},\end{aligned} \tag{3.12}$$

$$\boldsymbol{V}=\sum_{i=1}^{n}\boldsymbol{X_i}\boldsymbol{X_i}^{\boldsymbol{T}}\frac{1}{h}H'\left(\frac{Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_0}{h}\right)-nN\boldsymbol{\Sigma}^{-1}, \tag{3.13}$$

则有 (详细推导过程见附录):

$$\boldsymbol{V}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)=\boldsymbol{U}, \tag{3.14}$$

从而

$$\widehat{\boldsymbol{\beta}}_h=\widetilde{\boldsymbol{\beta}}+\boldsymbol{V}^{-1}\boldsymbol{U}. \tag{3.15}$$

注意到 $\boldsymbol{U}$, $\boldsymbol{V}$ 都是关于 $\widehat{\boldsymbol{\beta}}_h$ 的量，于是我们考虑 plug-in 估计，每一步在等式右边将 $\widehat{\boldsymbol{\beta}}_h$ 代入，计算新的 $\widehat{\boldsymbol{\beta}}_h$，一般的，在每一步迭代中，将上一步得到的估计值记作 $\widehat{\boldsymbol{\beta}}_0$ 得到 $\widehat{\boldsymbol{\beta}}_h$ 的迭代公式:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_h =&\widetilde{\boldsymbol{\beta}}+\left\{\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\boldsymbol{T}}\frac{1}{h}H'\Big(\frac{Y_i-\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_h}{h}\Big)-nN\boldsymbol{\Sigma}^{-\mathbf{1}}\right\}^{-1}\\&\times\left\{\sum_{i=1}^{n}\boldsymbol{X}_i\Big[H\Big(\frac{Y_i-\boldsymbol{X}_i^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_0}{h}\Big)+\tau-1+\frac{Y_i}{h}H'\Big(\frac{Y_i-\boldsymbol{X}_i^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_0}{h}\Big)\Big]\right.\\&\left.-\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\boldsymbol{T}}\frac{1}{h}H'\Big(\frac{Y_i-\boldsymbol{X}_i^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_0}{h}\Big)\widetilde{\boldsymbol{\beta}}\right\}.\end{aligned} \tag{3.16}$$

具体的算法流程如算法1所示。

### 3.2.2　回归系数方差估计公式

将 Jin et al.(2001)[42]提出的的针对目标函数的扰动重抽样方法应用在本文的回归系数方差估计上。分位数回归的估计问题实际上就是一个核函数带参数的 U 统计量求最小值问题，在分位数回归问题当中，U 统计量的阶数为 $K=1$，按照 Jin et al.(2001)[42]的要求，扰动变量 $V$ 的选取应满足均值为 $\mu$，方差为 $K^2\mu=\mu$，所以本文选择 Gamma(1,1) 作为扰动变量的分布。

值得注意的是，因为我们考虑了外部信息的随机性，所以回归系数估计值的 $\widehat{\boldsymbol{\beta}}$ 的随机性不只来自总体 $(\boldsymbol{X},Y)$，也有一部分随机性来自外部信息 $\widetilde{\boldsymbol{\beta}}$ 本身，因而辅助信息的随机性不只体现在引入的 $\boldsymbol{\Sigma}$ 当中，所以我们同样要对辅助信息进行扰动估计其随机性。由于这里假设辅助信息就是通过普通分位数回归算法得到的，所以在扰动辅助信息时会比扰动目标函数的求解过程要简单一些。记 $\xi_i^{ex}$, $i=1,...,n$ 为独立同分布于 Gamma(1,1)

的一列扰动变量，则添加了扰动变量之后，辅助信息的解为：

$$\widetilde{\boldsymbol{\beta}}^{*}=\underset{\boldsymbol{\beta}\in\mathbf{R}^{p+1}}{\arg\min}\sum_{i=1}^{n}\xi_{i}^{ex}\rho_{\tau}(Y_{i}-\boldsymbol{X_{i}}^{\boldsymbol{T}}\boldsymbol{\beta}). \tag{3.17}$$

记 $\xi_{i}^{in},\ i=1,...,n$ 为独立同分布于 Gamma(1,1) 的一列扰动变量，则添加了扰动变量之后，目标函数变为：

$$\mathcal{L}\left(\boldsymbol{\beta}\right)=\frac{1}{n}\sum_{i=1}^{n}\xi_{i}^{in}\left\{\rho_{\tau}\left(Y_{i}-X_{i}^{T}\boldsymbol{\beta}\right)-N\left(\widetilde{\boldsymbol{\beta}}^{*}-\boldsymbol{\beta}\right)^{T}\boldsymbol{\Sigma}^{-1}\left(\widetilde{\boldsymbol{\beta}}^{*}-\boldsymbol{\beta}\right)/2\right\}. \tag{3.18}$$

将示性函数用光滑函数 $H\left(\cdot\right)$ 近似代替后，由极值的一阶条件可以求出扰动重抽样下 (3.18) 回归系数的解：

记

$$\begin{aligned}\boldsymbol{U}^{*}=&\sum_{i=1}^{n}\xi_{i}^{in}\left[\boldsymbol{X_{i}}\left\{H\Big(\frac{Y_{i}-\boldsymbol{X_{i}}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_{0}}{h}\Big)+\tau-1+\frac{Y_{i}}{h}H'\Big(\frac{Y_{i}-\boldsymbol{X_{i}}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_{0}}{h}\Big)\right\}\right.\\&\left.-\sum_{i=1}^{n}\boldsymbol{X_{i}}\boldsymbol{X_{i}}^{\boldsymbol{T}}\frac{1}{h}H'\Big(\frac{Y_{i}-\boldsymbol{X_{i}}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_{0}}{h}\Big)\widetilde{\boldsymbol{\beta}}\right],\end{aligned} \tag{3.19}$$

$$\boldsymbol{V}^{*}=\sum_{i=1}^{n}\xi_{i}^{in}\left\{\boldsymbol{X_{i}}\boldsymbol{X_{i}}^{\boldsymbol{T}}\frac{1}{h}H'\Big(\frac{Y_{i}-\boldsymbol{X_{i}}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_{0}}{h}\Big)-N\boldsymbol{\Sigma}^{-1}\right\}, \tag{3.20}$$

则扰动重抽样样本下，回归系数的估计值为：

$$\widehat{\boldsymbol{\beta}}_{h}^{*}=\widetilde{\boldsymbol{\beta}}^{*}+\left(\boldsymbol{V}^{*}\right)^{-1}\boldsymbol{U}^{*}. \tag{3.21}$$

扰动重抽样估计系数方差的算法见算法2.

**算法 1** 带辅助信息的分位数回归算法

**Input:**

内部研究数据 $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{n}$；
辅助信息协方差矩阵 $\boldsymbol{\Sigma}$；
辅助信息 $\widetilde{\boldsymbol{\beta}}$；
分位数水平 $\tau$；
最大迭代数 $q$；
光滑函数 $H$；
窗宽 $h$；

**Output:** 回归系数估计值 $\widehat{\boldsymbol{\beta}}_h$

**for** $g = 1, 2, ..., q$ **do** //$g$ 表示迭代次数
　**if** $g = 1$ **then**
　　初始化 $\widehat{\boldsymbol{\beta}}_0$，采用最小二乘估计：

$$\widehat{\boldsymbol{\beta}}_0 = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}\right)^2$$

　**else**
　　将初始值 $\widehat{\boldsymbol{\beta}}_0$ 设置为上一步迭代得到的估计值 $\widehat{\boldsymbol{\beta}}^{(g-1)}$
　　按照公式3.15计算新的估计值 $\widehat{\boldsymbol{\beta}}^{(g)}$：

$$\widehat{\boldsymbol{\beta}}_h = \widetilde{\boldsymbol{\beta}} + \boldsymbol{V}^{-1}\boldsymbol{U}$$

　**end if**
　**if** $\|\widehat{\boldsymbol{\beta}}^{(g-1)} - \widehat{\boldsymbol{\beta}}^{(g)}\|_2 < 10^{-8}$ **then**
　　**break**
　**end if**
**end for**
**return** 回归系数估计值 $\widehat{\boldsymbol{\beta}}_h$

### 3.3　数值模拟

在本节，我们将进行数值模拟验证所提出算法的有效性，在实验中，光滑函数取 (3.8)，窗宽 $h$ 取为 $0.2n^{-0.2}$，辅助信息的协方差矩阵取为 $p+1$ 维单位阵，从如下线性模型中产生数据：

$$Y_i = \boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta} + \epsilon_i,\ i = 1, 2, ..., n, \tag{3.22}$$

其中 $\boldsymbol{X_i}^{\boldsymbol{T}} = (1, X_{i1}, ..., X_{ip})^T \in \mathbf{R}^{p+1}$ 是解释变量组成的向量。这里 $(X_{i1}, ..., X_{ip})^T$ 服从多元均匀分布 Uniform([0,1]$^p$)，且具有相关性 Corr$(X_{ij}, X_{ik}) = 0.5^{|j-k|}$, $1 \le j \ne k \le p$，本文生成具有指定相关性的多维均匀分布的方法参照 Falk(1999)[46]。模型 (3.22) 的回归系数设置为 $\boldsymbol{\beta} = \mathbf{1}_{p+1}$，误差 $\epsilon_i$ 分别服从如下分布：

1. 同方差正态：$\epsilon_i \sim N(0, 1)$,
2. 异方差正态：$\epsilon_i \sim N(0, (1 + 0.3X_{i1})^2)$,
3. 指数分布：$\epsilon_i \sim Exp(1)$.

本文中选定 $\tau = 0.25, 0.75, 0.9$ 三个分位数水平进行实验，对每一个选定的分位数水平，计算分位数回归系数的真实值 $\boldsymbol{\beta}(\tau)$，使得假设 $P(\epsilon_i \le 0 | \boldsymbol{X_i}) = \tau$ 成立，在各个误差假设下，回归系数真值如下：

1. 同方差正态：$\boldsymbol{\beta}(\tau) = \boldsymbol{\beta} + \Phi^{-1}(\tau)e_1$,
2. 异方差正态：$\boldsymbol{\beta}(\tau) = \boldsymbol{\beta} + \Phi^{-1}(\tau)e_1 + 0.3\Phi^{-1}(\tau)e_2$,
3. 指数分布：$\boldsymbol{\beta}(\tau) = \boldsymbol{\beta} + F_{exp}^{-1}e_1$.

这里 $\Phi$ 和 $F_{exp}$ 分别表示标准正态分布和比率参数为 1 的指数分布的累积分布函数，$\boldsymbol{e_i}$ 是一个 $p+1$ 维向量，仅第 $i$ 个元素为 1, 其余元素全为 0。窗宽 $h = 0.2n^{-0.2}$，光滑函数 $H(\cdot)$ 的选择为：

$$H(v) = \begin{cases} 0 & \text{if } v \le -1, \\ \frac{1}{2} + \frac{15}{16}\left(v - \frac{2}{3}v^3 + \frac{1}{5}v^5\right) & \text{if } |v| < 1, \\ 1 & \text{if } v \ge 1. \end{cases} \tag{3.23}$$

显然它二阶可导且二阶导数有限，且窗宽 $h = o(1)$.

**算法 2** 扰动重抽样估计系数方差

**Input:**

内部研究数据 $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$;
辅助信息协方差矩阵 $\boldsymbol{\Sigma}$;
辅助信息 $\widetilde{\boldsymbol{\beta}}$;
分位数水平 $\tau$;
最大迭代数 $q$;
光滑函数 $H$;
窗宽 $h$;
外部扰动变量 $\xi_i^{ex}$;
内部扰动变量 $\xi_i^{in}$;
重抽样次数 $B$;

**Output:** 标准差向量 $\boldsymbol{SD} = (SD_0, ..., SD_p)^T$

**for** $b = 1, 2, ..., B$ **do**
　产生外部扰动变量 $\xi_i^{ex},\ i = 1, ..., N$
　产生内部扰动变量 $\xi_i^{in},\ i = 1, ..., n$
　求解辅助信息的扰动估计 $\widetilde{\boldsymbol{\beta}}^*$:

$$\widetilde{\boldsymbol{\beta}}^* = \arg\min_{\boldsymbol{\beta} \in \mathbf{R}^{p+1}} \sum_{i=1}^{n} \xi_i^{ex} \rho_\tau (Y_i - \boldsymbol{X_i}^{\boldsymbol{T}} \boldsymbol{\beta})$$

　**for** $g = 1, 2, ..., q$ **do**
　　**if** $g = 1$ **then**
　　初始化 $\widehat{\boldsymbol{\beta}}_0$，采用最小二乘估计:

$$\widehat{\boldsymbol{\beta}}_0 = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \boldsymbol{X_i}^{\boldsymbol{T}} \boldsymbol{\beta}\right)^2$$

　　**else**
　　将初始值 $\widehat{\boldsymbol{\beta}}_0$ 设置为上一步迭代得到的估计值 $\widehat{\boldsymbol{\beta}}^{(g-1)}$
　　按照公式 (3.21) 计算新的估计值 $\widehat{\boldsymbol{\beta}}^{(g)}$:

$$\widehat{\boldsymbol{\beta}}_h = \widetilde{\boldsymbol{\beta}} + \boldsymbol{V}^{-1}\boldsymbol{U}$$

　　**end if**
　　**if** $\|\widehat{\boldsymbol{\beta}}^{(g-1)} - \widehat{\boldsymbol{\beta}}^{(g)}\|_2 < 10^{-8}$ **then**
　　**break**
　　**end if**
　**end for** // 得到回归系数的第 $b$ 个扰动估计 $\widehat{\boldsymbol{\beta}}_h^{(b)}$
**end for** // 得到 $B$ 个扰动估计
根据 $B$ 个扰动估计系数，对每一个分量 $\widehat{\boldsymbol{\beta}}_{hi}, i = 0, ..., p$，计算其样本标准差:

$$SD_i = \frac{1}{B-1} \sum_{b=1}^{B} \left(\widehat{\boldsymbol{\beta}}_{hi}^{(b)} - \bar{\boldsymbol{\beta}}_{hi}^*\right)$$

其中 $\bar{\boldsymbol{\beta}}_{hi}^* = \frac{1}{B}\sum_{b=1}^{B} \widehat{\boldsymbol{\beta}}_{hi}^{(b)}$
**return** 标准差向量 $\boldsymbol{SD} = (SD_0, ..., SD_p)^T$

在本文中，使用对目标函数进行扰动重抽样的方法估计回归参数 $\boldsymbol{\beta}$ 的标准差，进而计算置信区间和覆盖率，我们将步骤概括如下：

1. 对目标函数进行 $B$ 次扰动重抽样，获得 $B$ 个回归系数的估计值 $\left\{\boldsymbol{\beta}_j^*\right\}_{j=1}^B$

2. 计算 $B$ 个扰动重抽样样本的协方差矩阵，提取对角线元素作为回归系数 $\widehat{\boldsymbol{\beta}}_h$ 的估计值，$\boldsymbol{\beta}_j^* = \left(\beta_{0j}^*, \beta_{1j}^*, ..., \beta_{pj}^*\right)^T$，记 $\bar{\beta}_i^* = \frac{1}{B}\sum_{j=1}^B \beta_{ij}^*$ 是第 $i$ 个回归系数的 Bootstrap 样本均值，则第 $i$ 个回归系数的标准差 $SD_i$ 为：

$$SD_i = \sqrt{\frac{1}{B-1}\sum_{j=1}^{B}\left(\beta_{ij}^* - \bar{\beta}_i^*\right)^2}. \tag{3.24}$$

3. 对第 $i$ 个回归系数 $\beta_i$, $i = 0, ..., p$，构造其 $(1-\alpha)\%$ 置信区间，$z_{\frac{\alpha}{2}}$ 表示标准正态分布的下 $\frac{\alpha}{2}$ 分位数，$\widehat{\beta}_i$ 为 $\beta_i$ 的估计值：

$$\left[\widehat{\beta}_i - z_{\frac{\alpha}{2}}SD_i,\ \widehat{\beta}_i + z_{1-\frac{\alpha}{2}}SD_i\right]. \tag{3.25}$$

4. 总共进行 $N_s$ 次模拟，每一次都会获得一个回归系数的估计值 $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_0, ..., \widehat{\beta}_p\right)^T$ 和一组标准差向量 $\boldsymbol{SD} = (SD_0, ..., SD_p)$，对每一个回归系数构造上述置信区间，若真值 $\boldsymbol{\beta}_i$ 落在 $\widehat{\boldsymbol{\beta}}_i$ 的置信区间内，则记 1 次。最终对每一个回归系数，得到在 $N_s$ 次模拟中其真值落入置信区间的次数 $N_i$，由此可计算出每一个回归系数的经验覆盖率：

$$Cp_i = \frac{N_i}{N_s}. \tag{3.26}$$

若估计算法效率比较高，对真值和标准差估计的都比较准确，那么由上式计算的出的经验覆盖率 $Cp_i$ 应当接近 $(1-\alpha)\%$

进行 500 次模拟，并将结果与不带辅助信息的分位数回归相比较，同样使用扰动重抽样的方法估计不带辅助信息的分位数回归系数标准差。表3-1展示了在同方差标准正态误差下，在 0.25, 0.75 和 0.9 三个分位数水平下，两种算法的经验覆盖率，估计偏差、样本标准差以及总体标准差，其中总体标准差由扰动重抽样方法估算得来。理想情况下表中的样本标准差 std 和扰动重抽样得到的总体标准差 SD 应当非常接近，表明使用扰

动重抽样估计回归系数的方差是有效的，且每个回归系数的 90% 和 95% 经验覆盖率应当接近 0.9 和 0.95，表明当样本量趋于无穷时，回归系数的分布是渐进正态的。表3-2展示了在异方差正态下，两种算法的结果，而表3-3展示了在参数为 1 的指数分布误差下，两种算法的结果，其中列的含义同表3-1.

**表 3-1**　同方差正态误差下两种算法的比较
Table 3-1　Comparison when homoscedastic

| Coefficient | 带辅助信息 | | | | | 不带辅助信息 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Coverage Rate(90%) | Coverage Rate(95%) | Bias ($\times 10^{-2}$) | std | SD | Coverage Rate(90%) | Coverage Rate(95%) | Bias ($\times 10^{-2}$) | std | SD |
| $\tau = 0.25$ | | | | | | | | | | |
| $\beta_0$ | 0.890 | 0.922 | -0.71 | 0.111 | 0.109 | 0.890 | 0.940 | 0.00 | 0.107 | 0.108 |
| $\beta_1$ | 0.882 | 0.936 | -0.49 | 0.180 | 0.176 | 0.896 | 0.944 | -0.42 | 0.171 | 0.176 |
| $\beta_2$ | 0.896 | 0.926 | -0.49 | 0.180 | 0.176 | 0.884 | 0.944 | 0.65 | 0.192 | 0.195 |
| $\beta_3$ | 0.900 | 0.954 | 0.27 | 0.168 | 0.176 | 0.880 | 0.934 | 0.06 | 0.180 | 0.175 |
| $\tau = 0.75$ | | | | | | | | | | |
| $\beta_0$ | 0.876 | 0.934 | -0.03 | 0.108 | 0.109 | 0.910 | 0.956 | 0.00 | 0.109 | 0.110 |
| $\beta_1$ | 0.886 | 0.948 | -0.05 | 0.173 | 0.176 | 0.900 | 0.946 | 0.05 | 0.178 | 0.175 |
| $\beta_2$ | 0.910 | 0.952 | -0.04 | 0.186 | 0.195 | 0.890 | 0.944 | -0.08 | 0.198 | 0.195 |
| $\beta_3$ | 0.904 | 0.950 | -0.13 | 0.169 | 0.170 | 0.908 | 0.960 | 0.00 | 0.168 | 0.175 |
| $\tau = 0.9$ | | | | | | | | | | |
| $\beta_0$ | 0.912 | 0.960 | 0.02 | 0.131 | 0.138 | 0.896 | 0.942 | -0.02 | 0.134 | 0.137 |
| $\beta_1$ | 0.902 | 0.940 | 0.00 | 0.210 | 0.221 | 0.898 | 0.954 | 0.11 | 0.213 | 0.220 |
| $\beta_2$ | 0.894 | 0.942 | 0.01 | 0.239 | 0.245 | 0.900 | 0.946 | -0.03 | 0.235 | 0.245 |
| $\beta_3$ | 0.894 | 0.950 | -0.11 | 0.218 | 0.222 | 0.916 | 0.950 | -0.03 | 0.212 | 0.220 |

由表3-1可知，在同方差标准正态误差假设下，设定分位数水平为 $\tau = 0.25$ 时，不带辅助信息的算法对截距项估计得更好，其偏差和标准差都显著低于带辅助信息的情况；对于 $\beta_1$, $\beta_2$ 的估计两个算法的偏差和方差表现相似，但从经验覆盖率角度考虑，不带辅助信息的算法拥有更接近理论值的经验覆盖率，可以认为其表现更好；对于 $\beta_3$ 的估计，虽然不带辅助信息的算法比带辅助信息的算法得到的估计偏差更小，但是由于前者标准差也更大，所以导致在经验覆盖率上带辅助信息的算法表现更好。

在分位数水平为 $\tau = 0.75$ 时，两个算法在偏差上的表现都足够好，并且每个系数的估计方差也接近，从经验覆盖率上看，带辅助信息的算法的 90% 经验覆盖率的表现不如不带辅助信息的算法，但总体相差不大，依然与理论值相近。

在分位数水平为 $\tau = 0.9$ 时，两个算法的表现都很好，而带辅助信息的算法无论是从偏差还是经验覆盖率方面都比不带辅助信息的算法表现要好，带辅助信息的算法的 90% 和 95% 经验覆盖率都更接近理论值。纵向地看，本文提出的带随机性辅助信息的分位数回归算法随着分位数水平 $\tau$ 增大，估计偏差呈现出减小的趋势，算法对于标准正态误差下对于条件分布 $F(Y \mid \boldsymbol{X})$ 左尾的拟合效果不如不带辅助信息的算法，偏差稍

大。但即便如此，由于表格所示偏差的单位是 $10^{-2}$，带辅助信息的算法估计已经足够接近真值，可以认为是无偏估计。

另外，无论是带辅助信息还是不带辅助信息的情况，在同方差标准正态误差假设下，表格中的 std 列与对应的 SD 列的数值都足够接近，也就是说本文所采用的对目标函数进行扰动重抽样估计参数协方差矩阵的办法是有效的，至少对协方差矩阵的对角元的估计表现是令人满意的。

**表 3-2**　异方差正态误差下两种算法的比较
Table 3-2　Comparison when heteroscedastic

| Coefficient | 带辅助信息 | | | | | 不带辅助信息 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Coverage Rate(90%) | Coverage Rate(95%) | Bias ($\times 10^{-2}$) | std | SD | Coverage Rate(90%) | Coverage Rate(95%) | Bias ($\times 10^{-2}$) | std | SD |
| $\tau = 0.25$ | | | | | | | | | | |
| $\beta_0$ | 0.902 | 0.940 | -0.08 | 0.118 | 0.120 | 0.922 | 0.968 | -0.04 | 0.112 | 0.120 |
| $\beta_1$ | 0.888 | 0.946 | 0.23 | 0.196 | 0.201 | 0.904 | 0.960 | -0.05 | 0.198 | 0.203 |
| $\beta_2$ | 0.884 | 0.940 | -0.02 | 0.223 | 0.222 | 0.892 | 0.932 | 0.07 | 0.226 | 0.224 |
| $\beta_3$ | 0.892 | 0.940 | 0.00 | 0.205 | 0.120 | 0.906 | 0.950 | 0.12 | 0.191 | 0.202 |
| $\tau = 0.75$ | | | | | | | | | | |
| $\beta_0$ | 0.900 | 0.956 | 0.05 | 0.113 | 0.119 | 0.896 | 0.952 | 0.04 | 0.116 | 0.121 |
| $\beta_1$ | 0.912 | 0.968 | 0.03 | 0.194 | 0.200 | 0.908 | 0.954 | 0.05 | 0.190 | 0.200 |
| $\beta_2$ | 0.912 | 0.960 | -0.11 | 0.206 | 0.224 | 0.894 | 0.940 | -0.05 | 0.227 | 0.223 |
| $\beta_3$ | 0.916 | 0.954 | -0.05 | 0.195 | 0.201 | 0.902 | 0.954 | -0.04 | 0.195 | 0.200 |
| $\tau = 0.9$ | | | | | | | | | | |
| $\beta_0$ | 0.888 | 0.934 | -0.03 | 0.152 | 0.152 | 0.902 | 0.946 | 0.03 | 0.139 | 0.149 |
| $\beta_1$ | 0.902 | 0.948 | -0.05 | 0.248 | 0.253 | 0.922 | 0.954 | -0.07 | 0.236 | 0.253 |
| $\beta_2$ | 0.900 | 0.936 | -0.07 | 0.269 | 0.280 | 0.904 | 0.940 | -0.08 | 0.279 | 0.283 |
| $\beta_3$ | 0.882 | 0.932 | -0.05 | 0.253 | 0.253 | 0.898 | 0.948 | -0.06 | 0.247 | 0.251 |

由表3-2可知，在异方差正态误差假定下，带辅助信息的算法当 $\tau = 0.75$ 和 $\tau = 0.9$ 时，对 $\beta_1$ 和 $\beta_2$ 的估计的偏差分别稍大，但仍然是 $10^{-3}$ 量级的误差，总的来说，从经验覆盖率上看，两个算法的表现相似，都比较接近理论覆盖率。并且 std 和 SD 列的数值也相近，表明在异方差正态误差情况下，扰动目标函数的协方差估计算法仍然是有效的。

由表3-3可知，在比率参数为 1 的指数分布误差假设下，在给定的三个分位数水平，带辅助信息和不带辅助信息的算法在偏差以及方差上的表现都很好，偏差和方差都比较小。从经验覆盖率的角度考虑，带辅助信息的算法在三个分位数水平上的 95% 经验覆盖率比不带辅助信息的算法更接近理论值，表明本文提出的考虑辅助信息随机性的分位数回归算法有比较好的渐进性质，而在 90% 经验覆盖率上，两个算法的表现相似。

另外，注意到表3-3中的 std 列与对应的 SD 列同样非常接近，这表明在参数为 1 的指数误差假设下，本文所使用的针对目标函数进行扰动重抽样估计参数标准差的方法具有很好的表现；此外，纵向对比三个表格，可以看出本文提出的分位数回归算法对误差

假设是相对稳健的，并且本文所采用的方差估计方法对于误差假设也是较为稳健的。

**表 3-3**　指数误差下两种算法的比较

Table 3-3　Comparison under exponential with rate 1

| Coefficient | 带辅助信息 | | | | | 不带辅助信息 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Coverage Rate(90%) | Coverage Rate(95%) | Bias ($\times 10^{-2}$) | std | SD | Coverage Rate(90%) | Coverage Rate(95%) | Bias ($\times 10^{-2}$) | std | SD |
| $\tau = 0.25$ | | | | | | | | | | |
| $\beta_0$ | 0.898 | 0.952 | 0.00 | 0.044 | 0.046 | 0.884 | 0.940 | 0.42 | 0.046 | 0.046 |
| $\beta_1$ | 0.918 | 0.954 | 0.11 | 0.071 | 0.074 | 0.898 | 0.930 | 0.07 | 0.072 | 0.074 |
| $\beta_2$ | 0.896 | 0.942 | -0.01 | 0.083 | 0.083 | 0.912 | 0.940 | 0.00 | 0.078 | 0.082 |
| $\beta_3$ | 0.902 | 0.964 | 0.00 | 0.072 | 0.074 | 0.872 | 0.928 | 0.04 | 0.075 | 0.073 |
| $\tau = 0.75$ | | | | | | | | | | |
| $\beta_0$ | 0.908 | 0.960 | -0.07 | 0.129 | 0.138 | 0.906 | 0.956 | -0.01 | 0.134 | 0.138 |
| $\beta_1$ | 0.906 | 0.952 | 0.11 | 0.217 | 0.225 | 0.918 | 0.946 | 0.01 | 0.211 | 0.223 |
| $\beta_2$ | 0.880 | 0.960 | -0.09 | 0.241 | 0.247 | 0.920 | 0.954 | 0.08 | 0.238 | 0.246 |
| $\beta_3$ | 0.878 | 0.934 | 0.12 | 0.224 | 0.224 | 0.936 | 0.962 | 0.00 | 0.207 | 0.223 |
| $\tau = 0.9$ | | | | | | | | | | |
| $\beta_0$ | 0.924 | 0.950 | -0.09 | 0.125 | 0.140 | 0.906 | 0.954 | -0.02 | 0.128 | 0.137 |
| $\beta_1$ | 0.916 | 0.952 | -0.09 | 0.214 | 0.222 | 0.914 | 0.956 | 0.00 | 0.211 | 0.221 |
| $\beta_2$ | 0.898 | 0.954 | 0.00 | 0.240 | 0.246 | 0.894 | 0.944 | 0.01 | 0.244 | 0.245 |
| $\beta_3$ | 0.902 | 0.952 | 0.00 | 0.213 | 0.223 | 0.898 | 0.934 | 0.02 | 0.222 | 0.222 |

# 第四章　总结与展望

本文在分位数回归中引入了辅助信息，并考虑了辅助信息的随机性，提出了一种新的估计回归系数的方法，并使用直接扰动目标函数的重抽样方法估计回归系数的方差。结果表明，新提出的算法由于引入了更多的信息，并考虑了辅助信息的随机性，提升了算法的表现，相比不带辅助信息的分位数回归算法，得到的结果有更小的偏差和更低的方差，并且具有更接近理论值的覆盖率，并且直接扰动目标函数的扰动重抽样方法对回归系数的方差估计得较好。

这表明新提出的分位数回归算法不仅在理论上具有创新性，并且在实际应用当中也可能有比较明显的性能提升。通过引入辅助信息并考虑其随机性，算法能够更全面地捕捉数据的内在结构，从而提供更为精确和可靠的回归系数估计。此外，直接扰动目标函数的重抽样方法在估计回归系数方差方面表现出了较高的准确性，这进一步增强了我们方法的可靠性。

我们的研究结果还表明，新算法在处理具有辅助信息的数据集时，能够有效地减小估计偏差，降低方差，并提高置信区间的覆盖率。这些改进对于提升分位数回归模型的预测能力和解释力具有重要意义，尤其是在金融、经济和社会科学研究等领域，这些领域中辅助信息的获取相对容易，且对模型的准确性要求较高。

尽管我们的算法在估计效率方面表现出了一定的优越性，但仍然存在进一步改善的空间。首先就是辅助信息随机性的刻画，本文简单粗暴地选择了单位阵作为其协方差矩阵的估计，虽然结果表明，即使在最退化的情况下，算法依然有较好的表现，但是理论上若协方差矩阵的估计越靠近系数的真实协方差矩阵，算法的表现应当越好。在未来的工作中，可能会考虑如何对外部信息的协方差矩阵进行可靠的估计，期望进一步提升算法的性能。

其次，当前方法只考虑了样本量为 1000 的情况，尽管算法收敛速度很快，但当样本量呈指数级增长时，在大数据的背景下该算法是否仍然具有高效的计算效率是一个仍待考究的问题。未来的工作中，可能会考虑如何将算法推广到分布式的情况，并考虑如何才能尽量让通讯成本降低，同时保证算法的估计效率。

最后，本文针对引入辅助信息的分位数回归估计，只探索性地给出了迭代算法，并设计模拟实验验证了其有效性，在理论上还未有严格的证明，并且由于本文缺乏对算法更深层次的理论讨论，所以选用了扰动重抽样的方式估计系数的方差。在未来的工作中，应当对算法的理论性质进行推导，并推导出总体层面的方差公式，不依赖于重抽样的办法，降低对计算量的依赖。

综上所述，本文提出的分位数回归算法通过有效地利用辅助信息并考虑其随机性，显著提高了系数估计的准确性和可靠性。但仍有一些不足之处，留待未来的工作中继续改进。期望本文的工作能够为提高参数的估计效率提供新的视角，并启发未来的研究者在此领域进行更深入的探索和创新。

# 附录 A　附录

## A.1　公式推导

### A.1.1　带辅助信息的分位数回归迭代算法推导

目标是:

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} \rho_{\tau}\left(Y_i - \boldsymbol{X_i}^{\boldsymbol{T}} \boldsymbol{\beta}\right) - N\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{T} \boldsymbol{\Sigma^{-1}} \left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) /2,$$

用一光滑函数 $H(\frac{x}{h})$ 代替示性函数 $\mathbf{1}\{x \geq 0\}$, 得到分位数损失的近似:

$$\begin{aligned} K_h(x) &= x\left[H\left(\frac{x}{h}\right) + \tau - 1\right] \\ \frac{\partial K_h(x)}{\partial x} &= H\left(\frac{x}{h}\right) + \tau - 1 + \frac{x}{h} H'\left(\frac{x}{h}\right), \end{aligned}$$

目标变为:

$$\min_{\boldsymbol{\beta}} \boldsymbol{L}(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} K_h\left(Y_i - \boldsymbol{X_i}^{\boldsymbol{T}} \boldsymbol{\beta}\right) - N\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{T} \boldsymbol{\Sigma^{-1}} \left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) /2,$$

注意到:

$$\begin{aligned} &\frac{\partial K_h\left(Y_i - \boldsymbol{X_i}^{\boldsymbol{T}} \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} = \frac{\partial K_h\left(Y_i - \boldsymbol{X_i}^{\boldsymbol{T}} \boldsymbol{\beta}\right)}{\partial\left(Y_i - \boldsymbol{X_i}^{\boldsymbol{T}} \boldsymbol{\beta}\right)} \times \frac{\partial\left(Y_i - \boldsymbol{X_i}^{\boldsymbol{T}} \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} \\ &= (-\boldsymbol{X_i}) \left[H\left(\frac{Y_i - \boldsymbol{X_i}^{\boldsymbol{T}} \boldsymbol{\beta}}{h}\right) + \tau - 1 + \frac{Y_i - \boldsymbol{X_i}^{\boldsymbol{T}} \boldsymbol{\beta}}{h} H'\left(\frac{Y_i - \boldsymbol{X_i}^{\boldsymbol{T}} \boldsymbol{\beta}}{h}\right)\right], \end{aligned}$$

且

$$\frac{\partial\Big[N\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)^{T}\boldsymbol{\Sigma^{-1}}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)\Big]}{\partial\boldsymbol{\beta}}=\frac{\partial\Big[N\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)^{T}\boldsymbol{\Sigma^{-1}}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)\Big]}{\partial\Big(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\Big)}=\boldsymbol{\Sigma^{-1}}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right).$$

所以 $\frac{\partial \boldsymbol{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ 为:

$$\frac{1}{n}\sum_{i=1}^{n}(-\boldsymbol{X_i})\left[H\left(\frac{Y_i-\boldsymbol{X_i^T}\boldsymbol{\beta}}{h}\right)+\tau-1+\frac{Y_i-\boldsymbol{X_i^T}\boldsymbol{\beta}}{h}H'\left(\frac{Y_i-\boldsymbol{X_i^T}\boldsymbol{\beta}}{h}\right)\right]-N\boldsymbol{\Sigma^{-1}}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right).$$

令 $\frac{\partial \boldsymbol{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}=0$, 有

$$\sum_{i=1}^{n}\boldsymbol{X_i}\Big[H\left(\frac{Y_i-\boldsymbol{X_i^T}\boldsymbol{\beta}}{h}\right)+\tau-1+\frac{Y_i-\boldsymbol{X_i^T}\boldsymbol{\beta}}{h}H'\left(\frac{Y_i-\boldsymbol{X_i^T}\boldsymbol{\beta}}{h}\right)\Big]+nN\boldsymbol{\Sigma^{-1}}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)=0,$$

即

$$\begin{aligned}&\sum_{i=1}^{n}\boldsymbol{X_i}\Big\{H\Big(\frac{Y_i-\boldsymbol{X_i^T}\boldsymbol{\beta}}{h}\Big)+\tau-1+\frac{Y_i}{h}H'\Big(\frac{Y_i-\boldsymbol{X_i^T}\boldsymbol{\beta}}{h}\Big)\Big\}\\&-\sum_{i=1}^{n}\boldsymbol{X_i}\boldsymbol{X_i^T}\frac{1}{h}H'\Big(\frac{Y_i-\boldsymbol{X_i^T}\boldsymbol{\beta}}{h}\Big)\boldsymbol{\beta}+nN\boldsymbol{\Sigma^{-1}}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)=0,\end{aligned}$$

则有

$$\begin{aligned}&\sum_{i=1}^{n}\boldsymbol{X_i}\Big\{H\Big(\frac{Y_i-\boldsymbol{X_i^T}\boldsymbol{\beta}}{h}\Big)+\tau-1+\frac{Y_i}{h}H'\Big(\frac{Y_i-\boldsymbol{X_i^T}\boldsymbol{\beta}}{h}\Big)\Big\}-\sum_{i=1}^{n}\boldsymbol{X_i}\boldsymbol{X_i^T}\frac{1}{h}H'\Big(\frac{Y_i-\boldsymbol{X_i^T}\boldsymbol{\beta}}{h}\Big)\widetilde{\boldsymbol{\beta}}\\&+\sum_{i=1}^{n}\boldsymbol{X_i}\boldsymbol{X_i^T}\frac{1}{h}H'\Big(\frac{Y_i-\boldsymbol{X_i^T}\boldsymbol{\beta}}{h}\Big)\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)+nN\boldsymbol{\Sigma^{-1}}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)=0,\end{aligned}$$

记

$$\begin{aligned}\boldsymbol{U}=&\sum_{i=1}^{n}\boldsymbol{X_i}\left\{H\left(\frac{Y_i-\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_0}{h}\right)+\tau-1+\frac{Y_i}{h}H'\left(\frac{Y_i-\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_0}{h}\right)\right\}\\&-\sum_{i=1}^{n}\boldsymbol{X_i}\boldsymbol{X_i^T}\frac{1}{h}H'\Big(\frac{Y_i-\boldsymbol{X_i^T}\widehat{\boldsymbol{\beta}}_0}{h}\Big)\widetilde{\boldsymbol{\beta}},\\\boldsymbol{V}=&\sum_{i=1}^{n}\boldsymbol{X_i}\boldsymbol{X_i^T}\frac{1}{h}H'\left(\frac{Y_i-\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_h}{h}\right)-nN\boldsymbol{\Sigma^{-1}}.\end{aligned}$$

则有

$$\boldsymbol{V}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)=\boldsymbol{U},$$

从而

$$\widehat{\boldsymbol{\beta}}_h = \widetilde{\boldsymbol{\beta}} + \boldsymbol{V}^{-1}\boldsymbol{U},$$

即

$$\widehat{\boldsymbol{\beta}}_h = \widetilde{\boldsymbol{\beta}} + \left\{ \sum_{i=1}^{n} \boldsymbol{X_i}\boldsymbol{X_i^T}\frac{1}{h}H'\left(\frac{Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_h}{h}\right) - nN\boldsymbol{\Sigma^{-1}} \right\}^{-1} \times$$
$$\left[ \sum_{i=1}^{n} \boldsymbol{X_i}\Big\{ H\left(\frac{Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_0}{h}\right) + \tau - 1 + \frac{Y_i}{h}H'\left(\frac{Y_i - \boldsymbol{X_i^T}\widehat{\boldsymbol{\beta}}_0}{h}\right) \Big\} - \sum_{i=1}^{n} \boldsymbol{X_i}\boldsymbol{X_i}^{\boldsymbol{T}}\frac{1}{h}H'\Big(\frac{Y_i - \boldsymbol{X_i}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_0}{h}\Big)\widetilde{\boldsymbol{\beta}} \right].$$

### A.1.2　扰动重抽样迭代算法的推导

#### 辅助信息的扰动重抽样

辅助信息由如下公式得到：

$$\widetilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbf{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}).$$

加入扰动项，由如下公式得到辅助信息的扰动重抽样估计：

$$\widetilde{\boldsymbol{\beta}}^* = \underset{\boldsymbol{\beta}\in\mathbf{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \xi_i^{ex}\rho_\tau(Y_i - \boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}).$$

上式的求解可以通过 $R$ 中的 $quantreg$ 函数设置 $weights = \xi_i^{ex}$ 实现。

#### 回归系数的扰动重抽样

添加扰动后目标是

$$\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n} \xi_i^{in}\left\{ \rho_\tau\left(Y_i - \boldsymbol{X_i^T}\boldsymbol{\beta}\right) - N\left(\widetilde{\boldsymbol{\beta}}^* - \boldsymbol{\beta}\right)^T \boldsymbol{\Sigma}^{-1}\left(\widetilde{\boldsymbol{\beta}}^* - \boldsymbol{\beta}\right)/2 \right\}.$$

用一光滑函数 $H(\frac{x}{h})$ 代替示性函数 $\mathbf{1}\{x \geq 0\}$, 得到分位数损失的近似，目标变为

$$\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n} \xi_i^{in}\left\{ \rho_\tau\left(Y_i - \boldsymbol{X_i^T}\boldsymbol{\beta}\right) - N\left(\widetilde{\boldsymbol{\beta}}^* - \boldsymbol{\beta}\right)^T \boldsymbol{\Sigma}^{-1}\left(\widetilde{\boldsymbol{\beta}}^* - \boldsymbol{\beta}\right)/2 \right\}.$$

注意到

$$\begin{aligned}&\frac{\partial K_h\left(Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}=\frac{\partial K_h\left(Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}\right)}{\partial\left(Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}\right)}\times\frac{\partial\left(Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}\\&=(-\boldsymbol{X_i})\left[H\left(\frac{Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}}{h}\right)+\tau-1+\frac{Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}}{h}H'\left(\frac{Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}}{h}\right)\right],\end{aligned}$$

且

$$\frac{\partial\left[N\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}^*\right)^T\boldsymbol{\Sigma^{-1}}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}^*\right)\right]}{\partial\boldsymbol{\beta}}=\frac{\partial\left[N\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}^*\right)^T\boldsymbol{\Sigma^{-1}}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}^*\right)\right]}{\partial\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}^*\right)}=\boldsymbol{\Sigma^{-1}}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}^*\right).$$

所以令 $\frac{\partial\boldsymbol{L}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=0$, 有

$$\begin{aligned}&\sum_{i=1}^n\xi_i^{in}\left\{(-\boldsymbol{X_i})\left[H\left(\frac{Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}}{h}\right)+\tau-1+\frac{Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}}{h}H'\left(\frac{Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta}}{h}\right)\right]-N\boldsymbol{\Sigma^{-1}}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}^*\right)\right\}\\&=0.\end{aligned}$$

之后的推导同上，最后记

$$\begin{aligned}\boldsymbol{U}^*&=\sum_{i=1}^n\xi_i^{in}\left\{\boldsymbol{X_i}\left[H\left(\frac{Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_0}{h}\right)+\tau-1+\frac{Y_i}{h}H'\left(\frac{Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_0}{h}\right)\right]\right.\\&\left.-\sum_{i=1}^n\boldsymbol{X_i}\boldsymbol{X_i}^{\boldsymbol{T}}\frac{1}{h}H'\Big(\frac{Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_0}{h}\Big)\widetilde{\boldsymbol{\beta}}\right\},\\\boldsymbol{V}^*&=\sum_{i=1}^n\xi_i^{in}\left[\boldsymbol{X_i}\boldsymbol{X_i}^{\boldsymbol{T}}\frac{1}{h}H'\left(\frac{Y_i-\boldsymbol{X_i}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_0}{h}\right)-N\boldsymbol{\Sigma}^{-1}\right].\end{aligned}$$

得到回归系数的扰动重抽样样本估计值:

$$\widehat{\boldsymbol{\beta}}_h^*=\widetilde{\boldsymbol{\beta}}^*+\left(\boldsymbol{V}^*\right)^{-1}\boldsymbol{U}^*.$$

# 参考文献


[1] Koenker, R. Bassett, G. Regression Quantiles[J]. Econometrica, 1978, 46(1): 33-50.

[2] Buchinsky, M. Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression[J]. Econometrica, 1994, 62(2): 405-458.

[3] Baur, D. G. Financial contagion and the real economy[J]. Journal of Banking Finance, 2012, 36(10): 2680-2692.

[4] Wattanawongwan, S., Mues, C., Okhrati, R., Choudhry, T., So, M. C. Modelling credit card exposure at default using vine copula quantile regression[J]. European Journal of Operational Research, 2023, 311(1): 387-399.

[5] Taylor, J. W. A quantile regression approach to estimating the distribution of multiperiod returns[J]. Journal of Derivatives, 1999, 7(1): 64.

[6] Taylor, J. W. A quantile regression neural network approach to estimating the conditional density of multiperiod returns[J]. Journal of Forecasting, 2000, 19(4): 299-311.

[7] Hwang, C.-H. Support vector quantile regression for longitudinal data[J]. Journal of the Korean Data and Information Science Society, 2010, 21(2): 309-316.

[8] Koenker, R. Schorfheide, F. Quantile spline models for global temperature change[J]. Climatic Change, 1994, 28(4): 395-404.

[9] Pozarickij, A., Williams, C., Hysi, P. G., Guggenheim, J. A. Quantile regression analysis reveals widespread evidence for gene-environment or gene-gene interactions in myopia development[J]. Communications Biology, 2019, 2(1): 167.

[10] Zhang, H., Deng, L., Schiffman, M., Qin, J., Yu, K. Generalized integration model for improved statistical inference by leveraging external summary data[J]. Biometrika, 2020, 107(3): 689-703.

[11] Galton, F. Regression towards mediocrity in hereditary stature[J]. The Journal of the Anthropological Institute of Great Britain and Ireland, 1886, 15(1): 246-263.

[12] White, H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for het-

eroskedasticity[J]. Econometrica: Journal of the Econometric Society, 1980, 48(4): 817-838.

[13] Huber, P. J., Proceedings of the fifth Berkeley symposium on mathematical statistics and probability[M]. Berkeley, CA: University of California Press, 1967: 221-233.

[14] Carroll, R. J. Ruppert, D.Transformation and weighting in regression[M]. New York: Chapman, 2017: 161-165.

[15] Koenker, R. Bassett Jr, G. Robust tests for heteroscedasticity based on regression quantiles[J]. Econometrica: Journal of the Econometric Society, 1982, 50(1): 43-61.

[16] Bassett, G. W. Koenker, R. W. Strong consistency of regression quantiles and related empirical processes[J]. Econometric Theory, 1986, 2(2): 191-201.

[17] Buchinsky, M. Estimating the asymptotic covariance matrix for quantile regression models a Monte Carlo study[J]. Journal of Econometrics, 1995, 68(2): 303-338.

[18] Zhang, H., Deng, L., Schiffman, M., Qin, J., Yu, K. Generalized integration model for improved statistical inference by leveraging external summary data[J]. Biometrika, 2020, 107(3): 689-703.

[19] Lin, D.-Y. Zeng, D. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis[J]. Biometrika, 2010, 97(2): 321-332.

[20] Chen, J. Qin, J. Empirical likelihood estimation for finite populations and the effective usage of auxiliary information[J]. Biometrika, 1993, 80(1): 107-116.

[21] Chen, J. Sitter, R. A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys[J]. Statistica Sinica, 1999: 385-406.

[22] Imbens, G. W. Lancaster, T. Combining micro and macro data in microeconometric models[J]. The Review of Economic Studies, 1994, 61(4): 655-680.

[23] Chen, J, Sitter, R., Wu, C. Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys[J]. Biometrika, 2002, 89(1): 230-237.

[24] Cheng, W., Taylor, J. M., Gu, T., Tomlins, S. A., Mukherjee, B. Informing a risk prediction model for binary outcomes with external coefficient information[J]. Journal of the Royal Statistical Society Series C: Applied Statistics, 2019, 68(1): 121-139.

[25] Cheng, W., Taylor, J. M., Vokonas, P. S., Park, S. K., Mukherjee, B. Improving estimation and prediction in linear regression incorporating external information from an established reduced model[J]. Statistics in Medicine, 2018, 37(9): 1515-1530.

[26] Qin, J. Miscellanea. Combining parametric and empirical likelihoods[J]. Biometrika, 2000, 87(2): 484-490.

[27] Efron, B. Bootstrap Methods: Another Look at the Jackknife[J]. The Annals of Statistics, 1979, 7(1):

1-26.

[28] Efron, B. Nonparametric standard errors and confidence intervals[J]. Canadian Journal of Statistics, 1981, 9(2): 139-158.

[29] Efron, B. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods[J]. Biometrika, 1981, 68(3): 589-599.

[30] Efron, B. The jackknife, the bootstrap and other resampling plans[M]. Ohio: SIAM, 1982: 61-66.

[31] 欧变玲. 空间经济计量模型 Bootstrap Moran 检验有效性研究 [D]. 广东省: 华南理工大学, 2011.

[32] Wu, C.-F. J. Jackknife, bootstrap and other resampling methods in regression analysis[J]. The Annals of Statistics, 1986, 14(4): 1261-1295.

[33] Liu, R. Y. Bootstrap procedures under some non-iid models[J]. The Annals of Statistics, 1988, 16(4): 1696-1708.

[34] Mammen, E. Bootstrap and wild bootstrap for high dimensional linear models[J]. The Annals of Statistics, 1993, 21(1): 255-285.

[35] Freedman, D. A. Bootstrapping regression models[J]. The Annals of Statistics, 1981, 9(6): 1218-1228.

[36] Freedman, D. On bootstrapping two-stage least-squares estimates in stationary linear models[J]. The Annals of Statistics, 1984, 12(3): 827-842.

[37] Flachaire, E. A better way to bootstrap pairs[J]. Economics Letters, 1999, 64(3): 257-262.

[38] Politis, D. N. The impact of bootstrap methods on time series analysis[J]. Statistical Science, 2003, 18(2): 219-230.

[39] Gonçalves, S.  White, H. Maximum likelihood and the bootstrap for nonlinear dynamic models[J]. Journal of Econometrics, 2004, 119(1): 199-219.

[40] Horowitz, J. L. The bootstrap in econometrics[J]. Statistical Science, 2003, 18(2): 211-218.

[41] Hu, F.  Kalbfleisch, J. D. The estimating function bootstrap[J]. Canadian Journal of Statistics, 2000, 28(3): 449-481.

[42] Jin, Z., Ying, Z.,  Wei, L. A simple resampling method by perturbing the minimand[J]. Biometrika, 2001, 88(2): 381-390.

[43] Hoeffding, W. A Class of Statistics with Asymptotically Normal Distribution[J]. The Annals of Mathematical Statistics, 1948, 19(3): 293-325.

[44] White, H. Maximum likelihood estimation of misspecified models[J]. Econometrica: Journal of the Econometric Society, 1982, 50(1): 1-25.

[45] Chen, X., Liu, W., Zhang, Y. Quantile regression under memory constraint[J]. The Annals of Statistics, 2019, 47(6): 3244 -3273.

[46] Falk, M. A simple approach to the generation of uniformly distributed random variables with prescribed correlations[J]. Communications in Statistics-Simulation and Computation, 1999, 28(3): 785-791.

# 致　谢

时光荏苒，岁月如梭，不知不觉在师大的本科生涯已接近尾声。本人于 2019 年 9 月进入师大学习，于次年 4 月转入师大统计系，至今已四载有余，对师大更是有很深的感情。与统计系各位老师同学相处的时光让我受益匪浅，在本文从选题到完成的各个阶段，自始至终得到了导师周勇的悉心指导和帮助，并提供了许多宝贵的资料和建议，同组的苏瑾师姐也无私地为我解答困惑，在此由衷地感谢周老师和苏瑾师姐给予我无私的帮助！

在师大统计的四年，我有幸遇见了许多治学严谨、通情达理、亦师亦友的老师。感谢华东师范大学数学科学学院的吴瑞聪教授，带领我进入数学分析的大门，使我感受到数学的魅力，也为我今后四年的学习养成了良好的学习习惯；感谢华东师范大学统计学院的姚强老师，他风趣幽默的授课方式让我对数的世界有了更深的兴趣；感谢许忠好老师，他以同样引人入胜的方式指引我初探概率的世界；感谢刘玉坤老师，他教授的数理统计深入浅出，让我我迈入统计的大门；感谢周迎春老师，对我每一个不成熟的想法表示支持和肯定并提出可行的建议，极大地激发了我对科研的兴趣；感谢马慧娟老师，在封校期间，她除了认真严谨地完成教学任务之外，还会关心我们在学校的生活，照顾我们的情绪，在本文完成过程中，感谢马老师在百忙之中拨冗查看我的论文，提出了许多宝贵的修改意见；感谢同组的张澍一老师，在她的因果推断课上我感受到老师对学术认真严谨的态度，是我学习的榜样；感谢同组的史兴杰老师，在我迷茫于未来研究方向的选择时无私地为我提出建议，分析利弊，也教导了我一些生活和做人的道理，让我受益匪浅。另外，还要感谢姚强老师、许忠好老师和周迎春老师在夏令营推免时拨冗为我填写推荐信，他们的支持和指导给了彼时迷茫的我莫大的鼓励和帮助。给我留下深刻印象的老师远不止这些，由于篇幅限制，在此一并表示感谢和尊敬。

五年的本科生涯即将结束，感谢师大统计系对我的栽培；感谢无私为我提供指导的老师们；感谢一直站在我身后默默奉献的父母，感谢你们对我做的每个决定的全力支持；也感谢在师大相识相知的每一个伙伴，是你们的陪伴圆满了我五年的青春；最后，感谢我自己，在五年的时间里没有放弃探索和挑战自己，让我不断成长，当然，这离不开前面所有人的帮助。感谢这五年里为我提供过帮助的每一个人，感谢你们作为我前行路上的指明灯，也感谢遇到困难没有放弃挣扎的自己，感谢亦师亦友的每一个伙伴。

此文落笔完结之际，基本宣告人生一个阶段的结束，而同时也是另一个新阶段的开始。行文至此，回望五年的本科生涯，虽有坎坷，但不留遗憾。前路挑战重重却也充满精彩，在此祝愿自己，也祝愿身边的老师、朋友和家人们，无论走到人生的哪一个阶段，都能不忘初心，不留遗憾。