分类号 0212/32　　密级 公开
U D C 0004822　　编号 10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

论文题目 基于分层数据的分位数回归研究

研 究 生 姓 名: 杨小卜

指导教师姓名、职称: 郭精军 教授

学 科、专 业 名 称: 统计学 数理统计学

研 究 方 向: 复杂数据研究

提 交 日 期: 2023 年 5 月 30 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果. 尽我所知, 除了文中特别加以标注和致谢的地方外, 论文中不包含其他人已经发表或撰写过的研究成果. 与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意.

学位论文作者签名：杨小卜 签字日期：2023.5.30

导师签名：[signature] 签字日期：2023.5.30

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定, 同意（选择“同意”/“不同意”）以下事项:

1.学校有权保留本论文的复印件和磁盘, 允许论文被查阅和借阅, 可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库, 传播本学位论文的全部或部分内容.

学位论文作者签名：杨小卜 签字日期：2023.5.30

导师签名：[signature] 签字日期：2023.5.30

Quantile Regression Research Based on Stratified Data

Candidate: Xiaobo Yang

Supervisor: Jingjun Guo

摘 要

在日常生活中, 人们会接触到各式各样的、来源非常多元化的数据. 为了发掘与利用数据的潜在价值, 需要根据数据的特点构建各式各样的统计模型. 随着大数据时代的到来, 数据量的增加使得诸如分层数据这样具有复杂结构的数据出现. 目前关于分层数据的研究聚焦于模型的推广：从分层最小二乘回归模型到分层分位数回归模型、分层 logistic 回归模型. 虽然, 上述模型拓宽了数据的应用范围. 但是仍存在一些问题如下：上述模型中的数据是完整的, 没有考虑数据出现缺失这一更符合现实生活的情况；分层分位数回归模型因为分位数模型损失函数不可微的缺点, 导致估计精度降低, 分层分位数回归模型在惩罚参数求解过程中对两个惩罚参数分别使用不同的算法求解, 致使估计的效率降低.

因此, 本文针对上述两个问题研究具有分层特性数据的建模问题, 主要研究内容分为两部分：

(1) 研究响应变量随机缺失以及异方差情况下响应变量随机缺失的分层分位数回归模型的估计问题. 首先采用逆概率加权方法对随机缺失的响应变量进行处理并使用 LASSO 惩罚函数与满足 Oracle 性质的 ADALASSO 惩罚函数进行降维处理, 构建回归系数的估计, 证明参数估计量的渐近性质. 其次在异方差与响应变量随机缺失的假设条件下, 同样证明了参数估计量的渐近性质. 最后通过蒙特卡洛数值模拟与实际人体基因数据进行实例分析, 结果表明所提方法表现良好.

(2) 基于分层分位数回归模型损失函数不可微, 导致模型估计精度降低的缺点, 提出分层卷积平滑分位数回归模型. 首先使用核函数通过卷积方法平滑处理不可微的分位数损失函数, 使之成为具有良好性质的可微凸函数. 其次使用 LASSO 惩罚函数进行降维处理并在参数估计时, 使用多步 ADMM 算法进行求解. 同时模型的惩罚参数通过公式变换, 使用 CV(交叉验证)方法进行求解. 最后通过数值模拟与实例分析表明：所提方法在处理具有尖峰厚尾特性数据时相较于分层分位数回归模型、分层线性回归模型更为有效；估计效率明显优于分层分位数回归模型.

关键词：卷积　分层数据　分位数回归　缺失数据　变量选择

Abstract

In daily life, people come into contact with a wide variety of data sources. In order to explore and utilize the potential value of data, it is necessary to build various statistical models according to the characteristics of data. With the advent of the era of big data, the increase in the amount of data makes the appearance of data with complex structure such as stratified data. At present, the research on stratified data focuses on the promotion of models: from stratified least square regression model to stratified quantile regression model, stratified logistic regression model. However, the above model broadens the range of applications of the data. However, there are still some problems as follows: the data in the above model is complete and does not consider the missing data, which is more in line with real life; The loss function of stratified quantile regression model is not differentiable, so the accuracy of estimation is reduced. In the process of solving penalty parameters, the stratified quantile regression model uses different algorithms to solve the two penalty parameters, which reduces the efficiency of estimation. Therefore, this paper studies the modeling of data with stratified characteristics. The main research content is divided into two parts:

(1) Based on stratified data, the estimation problem of stratified quantile regression model with random loss of response variables and random loss of response variables under heteroscedasticity is studied.

Firstly, the inverse probability weighting method was used to process the randomly missing response variables and the LASSO penalty function was used to reduce the dimension. The regression coefficient was estimated to prove the asymptotic property of the parameter estimators. Secondly, the asymptotic properties of parameter estimators are also proved under the assumption of heteroscedasticity and random absence of response variables. Finally, Monte Carlo numerical simulation and actual human genetic data analysis show that the proposed method performs well.

(2) Based on the fact that the loss function of stratified quantile regression model is nondifferentiable, which leads to reduced accuracy of model estimation, a stratified convolutional smooth quantile regression model is proposed First, the kernel function is used to smooth the nondifferentiable quantile loss function by convolution method, so that it becomes a differentiable convex function with good properties Secondly, the LASSO penalty function is used for dimension reduction and the multi-step ADMM algorithm is used for parameter estimation. At the same time, the penalty parameters of the model are solved by CV (cross validation) method through formula transformation. Finally, numerical simulation and case analysis show that the proposed method is more effective than stratified quantile regression model and stratified linear regression model in processing data with peak thick tail characteristics; The estimation efficiency is obviously superior to the stratified quantile regression model.

Key words: Convolution; Stratified data; Quantile regression; Missing data; Variable selection

目 录

1 引言

1.1 研究背景

随着信息技术的快速发展与数据储存技术的进步, 数据的维度和复杂程度增加. 如何在庞杂的数据中寻找重要的信息、减少无用信息的干扰, 成为当今学者们的研究热点. 许多学者在数据降维方面进行了大量的研究, 如Tibshirani(1996)提出的 LASSO 惩罚函数、Fan 等(2001)与 Fan 等(2008)分别提出的 SCAD 惩罚函数与针对超高维数据的 SIS 方法、Rothman 等(2009)提出的广义门限法等. 上述方法可以剔除无关信息, 提高模型的精度.

人们在研究回归问题时会发现协变量具有分组结构. 分组数据存在于生活的各个领域, 例如, 在研究初诊后乳腺癌的复发问题上, 需要考虑的重要因素是患者不同的组织亚型, 因此需要按照不同的亚型分组后进行分析；在致病基因的变量筛选方面, 也需要按照不同的临床期等特点进行分组. 在经济金融领域中, 分组数据的应用更加广泛, 通常需按照地区, 领域等划分后再进行分析. 在大数据时代下, 分组的数据使用越来越广泛, 将数据进行分组的准则本身也包含许多重要的信息, 充分挖掘这些信息是一项有价值的工作.

如果忽略分组结构, 将会导致估计效率低下, 模型可解释性差等问题. 为了解决上述问题, Yuan 等(2006)提出了 Group LASSO 方法, Wang 等(2008)、Hu 等(2018)在 Yuan 等(2006)的基础上分别提出了具有 Oracle 性质的自适应 Group LASSO 与自适应 Group ELASTIC-NET 方法. 除了协变量具有分组结构, 响应变量同样也具有类似的结构, 称为分层结构. 目前关于分层数据的研究多聚焦于模型的推广, 在模型中考虑数据是完整的这一过于理想化的条件. 同时没有对模型的估计精度与估计速度进行进一步的优化. 因此, 本文从分层数据模型的不足出发, 对分层分位数回归模型进行改进及扩展研究.

1.2 文献综述与研究现状

1.2.1 分层数据研究现状

随着信息技术的发展与数据可获得性的提高，分层数据越来越频繁的出现在生物医学、经济学与社会学等领域. 为了对分层数据进行分析，Gertheiss 等(2010)提出了基于系数分解的线性回归模型，虽然该模型可以很好的处理具有分层结构的数据，但其效果依赖于基准层的选择. Ollier等(2017)提出了基于LASSO惩罚函数的分层线性回归模型，该模型可以自适应的选择基准层，同时模型的估计精度要更高. 虽然该模型可以很好的对分层数据的个性与共性回归系数进行估计，但线性最小二乘回归模型只有在误差项满足正态分布的情况下，才可以很好的反映响应变量与一个或多个协变量之间的线性关系. 在实际的金融与医学等数据中，数据常常呈现出“尖峰厚尾”、偏态等特征，如 Torrenté(2020)研究表明，超过 50%的基因数据不服从正态分布. 如果此时使用线性回归中模型，模型将不会具有无偏性等优良性质，产生较大的误差. 为此，Koneker 和 Bassett(1978)提出了分位数回归模型. 相较于线性最小二乘回归模型，分位数回归模型在误差项存在异常值与误差项不满足正态分布的情况下，仍可以得到较为稳健的估计. 同时，相较于线性回归模型，分位数回归模型更具有解释性.

因此分位数回归模型近些年来得到了学者们的广泛关注，Koneker 和 Bassett(1982)系统研究了分位数回归的假设检验与异方差的稳健性检验等问题，为分位数回归模型在实际问题中的应用打下了基础. 随后针对分位数回归模型的研究层出不穷；Koenker (2005)在给定协变量的情况下对整个响应变量的条件分布进行更全面的研究. 基于分位数回归模型的上述优点，刘栋等(2021)拓展了 Ollier 等(2017)的研究，将分层线性最小二乘回归模型拓展到了分层分位数回归模型. 该模型继承了分位数回归的优点，对有尖峰厚尾特征的数据具有稳健性.

1.2.2 缺失数据与异方差研究现状

分层线性最小二乘回归模型与分层分位数回归模型都假定数据是可以观察或者测量的，不存在缺失数据的情况，但在生物医学与金融等领域会因为各种原因，导致数据出现缺失的情况. 如果在估计过程中忽略缺失值，将会导致模型产生一定程度的偏差，降低模型的估计效率. 因此，对缺失数据下的模型进行相关分析研究具有相当重要的现实意义. Little 与 Rubin(2002)将缺失数据从缺失机制上分为 3 类，即随机缺失，完全随机缺失与非随机缺失. 近年来，随着计算机计

算的发展, 许多学者提出了可以有效处理缺失数据的几类方法. 例如：Horvitz 等(1952)提出的逆概率加权(IPW)方法， Rubin(2004) 提出的多重插补法等. Zhao 等(2016)利用插补法研究响应变量随机缺失下的部分线性分位数回归模型. Sherwood 等(2013), Sherwood(2016)使用 IPW 方法分别研究了协变量随机缺失的情况下, 加权分位数回归模型与高维部分线性分位数回归模型的变量选择问题. Han 等 (2019)提出结合 IPW 方法与插补法的框架去研究响应变量或协变量随机缺失时的分位数回归模型. Zhao 等(2015)在响应变量随机缺失的情况下研究部分线性分位数回归模型. Bai 等(2020)在协变量随机缺失与存在测试误差的基础上研究了高维分位数回归模型的变量选择问题.

上述文献是关于随机缺失的研究, 关于非随机缺失与完全随机缺失的文献如下：Zhao 等(2015)在因变量与部分协变量非随机缺失的情况下研究广义线性模型的识别与估计、于力超(2019)在非随机缺失机制下, 研究模型参数的估计方法并将其拓展. 刘庆丰等(2020)等针对广义线性模型, 在协变量随机缺失的情况下得出模型平均估计方法. 随机缺失相较于非随机缺失与完全随机缺失, 理论探讨与模拟算法较为完善, 可供参考的研究较为丰富, 故而文中主要针对变量随机缺失进行研究. 在实际应用线性最小二乘回归模型时, 经常会出现误差项为异方差的现象, 该现象违反了线性最小二乘回归模型的基本假设, 导致模型的估计精度降低. Christou 等(2018)研究指出, 分位数回归模型在处理误差项为异方差时比线性最小二乘回归模型更有优势. 近年来许多学者在研究分位数回归模型时也会考虑异方差的问题, 如：Zheng 等(2013)提出了基于 GARCH 模型的混合分位数回归模型、Fan 等(2019)改进了具有持续预测变量与条件异方差误差的预测分位数回归模型.

1.2.3 平滑分位数模型研究现状

虽然, 分位数回归模型在处理具有尖峰厚尾特性的数据时要优于线性回归模型, 但是, 由于分位数回归的损失函数是不可微且非平滑的分段函数, 会降低常见的诸如 LLA(2008)(2014)、QICD(2012)(2015)等算法的估计精度, 与理论结果存在一定的差异.

为了克服分位数损失函数不可微的缺点, Horowitz (1998)首次提出使用核函

数去平滑分段损失函数的方法，并证明平滑损失函数的估计渐近等同于标准分位数回归的估计. 后续的学者在此基础上对平滑分位数回归进行了更深入的研究，如 Wu 等(2015)研究了带有测量误差的删失平滑分位数回归的理论性质；Galvao 等(2016)研究了基于面板数据的平滑分位数回归. 虽然 Horowitz(1998)中提出的平滑方法相较于分段损失函数具有一定的优势，但该方法产生的平滑损失函数是一个非凸函数，导致最优化问题难以求解. 为此 He 等(2021)、Tan 等(2022)提出使用卷积方法来平滑处理分段损失函数，并证明卷积平滑损失函数是凸平滑损失函数，同时，通用蒙特卡洛模拟与实际数据验证，基于该方法的分位数回归模型的估计精度要高于基于核函数的平滑分位数回归模型与基于分段损失函数的分位数回归模型.

1.2.4 文献述评

通过对已有文献进行梳理可以发现，学者们对分层数据的建模做了大量研究，也取得了一些有意义的成果，为本文的研究提供了诸多借鉴与参考. 但通过深入分析，仍有以下值得改进的地方：

(1) 虽然关于分层数据回归模型的相关研究非常多，但是所考虑的均是数据完全不存在缺失的情况，关于数据缺失与异方差的分层数据回归模型的研究较为欠缺. 因此，可以在分层分位数回归模型中考虑响应变量缺失与异方差的情况，对分层分位数回归模型进行拓展研究.

(2) 现有的文献在使用分位数回归模型时，很少考虑因损失函数不可微而导致分位数回归模型精度下降这一问题. 且在估计惩罚参数时，估计方法过于复杂，导致估计效率降低. 因此，可以从以上的两个问题出发，从估计精度与估计效率方面对分层分位数回归模型进行改进.

1.3 研究意义

随着大数据时代的到来，数据集的维度显著提高，数据复杂程度增加. 如果忽略数据中的分层关系，会使模型存在偏差，降低估计的准确性以及模型的可解释性. 对具有复杂结构的数据进行建模一直是统计研究工作者的一个热门研究方向，也是一个热门课题. 因此，对分层数据进行建模与深入研究具有重要的理

论意义和现实意义.

(1) 理论意义

采相较于无缺失值的分层分位数模型, 存在缺失值的分层分位数模型更加符合实际的应用情况. 得到模型估计量的渐近正态性, 推广了分组分位数回归模型. 使用卷积方法平滑处理不可微的分层分位数回归模型的损失函数, 提升模型的估计精度与估计效率. 这丰富了分层数据分析的相关理论知识, 为以后的研究提供了思路.

(2) 现实意义

通过缺失数据情况下的分层数据的研究, 以及对模型估计精度与估计性能的改进, 可以让该模型在诸如生物医学、经济学等领域中得到更好的应用, 发掘出数据背后所蕴含的信息.

1.4 研究内容与结构安排

1.4.1 研究内容

本文研究分层数据的建模问题, 研究内容主要分为如下两个方面：

(1) 针对响应变量随机缺失和响应变量随机缺失且存在异方差的分层数据, 分别建立了高维分层分位数估计模型, 并在一定假设条件下证明估计的渐近性质. 使用 LASSO 惩罚函数对模型进行降维处理, 并通过蒙特卡洛模拟验证估计方法的有效性与精度. 最后将模型应用到 THP-1 人骨髓单核细胞白血病细胞中分化为巨噬细胞的数据来验证分层分位数回归模型相较于分层线性回归模型变量选择的准确性.

(2)采用卷积平滑方法去改进分层分位数回归模型, 得到分层卷积平滑分位数回归模型, 并使用 LASSO 惩罚函数对分层卷积平滑分位数回归模型进行降维处理, 模型的参数通过用多步 ADMM 算法进行求解. 通过数值模拟与实例分析对分层线性回归模型、分层分位数回归模型与分层卷积平滑分位数回归模型进行比较, 验证所提模型的估计精度与估计效率.

1.4.2 结构安排

第一章为引言. 首先简单介绍了分层数据的研究背景. 然后结合国内外文献介绍了分层数据、缺失数据、异方差和平滑分位数模型算法的研究现状. 最后陈述了文章的研究内容与创新之处.

第二章为预备知识. 介绍了 Oracle 性质、LASSO 惩罚函数、ADALASSO 惩罚函数与分层线性回归模型.

第三章为响应变量缺失下异方差分层分位数回归模型的估计. 首先采用逆概率加权方法对随机缺失的响应变量进行处理, 使用 LASSO 惩罚函数与 ADALASSO 惩罚函数进行降维处理, 并证明参数估计量的渐近性质. 其次在异方差与响应变量随机缺失的假设条件下, 也证明了参数估计量的渐近性质. 最后通过蒙特卡洛数值模拟与实例分析来验证所提方法的估计性能.

第四章为卷积平滑分位数回归在高维分层数据中的应用. 首先采用卷积方法对不可微的分层分位数回归模型的损失函数进行平滑处理, 得到卷积平滑分层分位数回归模型. 其次使用 LASSO 惩罚函数对模型进行降维处理并使用多步 ADMM 算法对参数进行求解. 最后通过蒙特卡洛数值模拟与实例分析来验证所提方法的估计性能.

第五章为研究的总结与展望. 总结了全文的研究工作和存在的不足之处, 并且对未来研究方向和内容做出简单的展望.

1.5 创新点

针对现有模型的不足, 本文的具体创新之处可以概括为如下两点：

第一, 扩展分层分位数回归模型的应用范围. 目前关于分层回归模型的研究都聚焦于模型的推广, 从分层线性回归模型到分层分位数回归模型、分层 logistic 回归模型等. 没有考虑响应变量与协变量是否存在缺失, 也没有将异方差的情况考虑在内. 因此, 本文在分层分位数回归模型的基础上进行扩展研究, 分别考虑响应变量随机缺失、响应变量随机缺失与异方差情况下分层分位数回归模型的估计问题、证明模型在上述情况下回归参数的渐近性质, 并使用 LASSO 惩罚函数与 ADALASSO 惩罚函数对模型进行降维处理. 通过蒙特卡洛模拟与实际数据验证所提估计的性能.

第二, 提升分层分位数回归模型的估计精度与估计效率. 虽然分层分位数回

归模型在处理具有尖峰厚尾特性数据时估计精度较高, 但其不可微的损失函数会导致估计精度降低. 因此首先, 使用卷积方法对不可微的损失函数进行平滑处理, 使其成为可微的损失函数, 减少算法结果与理论结果的差异性. 其次, 在估计惩罚参数时, 改进估计方法, 减少估计步骤, 从而提升模型的估计精度与估计效率. 最后，使用 LASSO 惩罚函数进行降维处理, 减少因维数引起维数灾难对估计精度的影响. 所提模型估计的性能通过蒙特卡洛模拟与实际数据进行验证.

2 预备知识

2.1 变量选择方法相关知识

2.2.1 Oracle 性质

为了评估模型估计的优劣性, Fan 等(2001)提出了如下的 Oracle 性质：

1 稀疏性：模型在参数估计时, 自动的将一些不重要的参数系数压缩至 0.

2 无偏性：参数估计应该是无偏的, 对于系数较大变量的参数估计要做到近似无偏.

3 连续性：为了避免模型的不稳定性, 参数估计与其所对应数据是连续的.

2.1.2 LASSO 惩罚函数

LASSO 惩罚函数是由 Tibshirani(1996)提出的一种变量选择方法, 其惩罚项如下：

$$p_{\lambda,a}(\beta_i)^{LASSO} = \lambda|\beta_i|.$$

LASSO 惩罚函数通过λ对参数进行压缩, λ越大, 压缩作用就越明显, 可以将部分参数压缩为 0, 从而减少模型中待估参数的数量. 但是其估计不满足 Oracle 性质中的无偏性, 在待估参数较大的情况下会产生偏差.

2.1.3 ADALASSO 惩罚函数

ADALASSO 惩罚函数是由 Zou(2006)提出的一种具有 Oracle 性质的惩罚函数, 其惩罚项如下：

$$p_{\lambda,a}(\beta_i)^{\mathrm{ADALASSO}} = \lambda_T \widehat{\omega}_i|\beta_i^*|.$$

其中$\widehat{\omega}_i = \left(\frac{1}{\hat{\beta}_i^{LASSO}}\right)^{\gamma}$, $\hat{\beta}_i^{LASSO}$是 LASSO 惩罚函数求解得出的参数, 可以将 ADALASSO 惩罚函数简单看成是两阶段 LASSO 惩罚函数. 如果γ=0, 那么 ADALASSO 惩罚函数将会变为 LASSO 惩罚函数. 可以发现如果$\hat{\beta}_i^{LASSO}$很大, 那么在β_i^*上的惩罚就很小, 避免较大的参数被过度惩罚, 从而使得估计量满足

Oracle 性质.

2.2 分层线性回归模型

2.2.1 符号

记总体样本容量为n, 样本分为K组, $K \geq 1$. 令n_k表示每组的样本容量, 满足$\sum_{k=1}^{K} n_k = n$. 令$Y = \left(\left(y^{(1)}\right)^T, \cdots, \left(y^{(K)}\right)^T\right)^T$, $y^{(k)}$表示第k组响应变量, $x^{(k)}$表示第k组协变量.

2.2.2 分层线性回归模型

假定$\left\{\left(y_i^{(k)}, x_i^{(k)}\right), i = 1, \cdots, n_k\right\}$表示第$k$组独立同分布的样本, 其中$y_i^{(k)}$为响应变量, $x_i^{(k)} = (x_{i1}^{(k)}, \cdots, x_{ip}^{(k)})$为协变量. 假设$\epsilon_i^{(k)} = y_i^{(k)} - x_i^{(k)}\beta^{(k)^*}$, 为了体现出分组的特性, 对$\beta^{(k)^*}$做如下分解：

$$\beta^{(k)^*} = \upsilon^* + \delta^{(k)^*}, \tag{2.1}$$

其中$\upsilon^* = (v_1^*, \cdots, v_p^*)$表示不对样本进行分层时的共性回归系数, $\delta^{(k)^*} = (\delta_1^{(k)^*}, \cdots, \delta_p^{(k)^*})$表示第$k$组的个性回归系数, 该分解方法的核心思想为: 分层后的回归系数是共性回归系数与个性回归系数的线性组合.

对于(2.1)式中的待估参数$(\upsilon, \delta^{(1)}, \cdots \delta^{(K)})$, 可以通过求解如下目标函数进行估计：

$$(\hat{\upsilon}, \hat{\delta}^{(1)}, \cdots \hat{\delta}^{(K)}) \in \underset{\upsilon, \delta^{(1)}, \cdots \delta^{(K)}}{argmin} \left\{\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left\|y_i^{(k)} - x_i^{(k)}\left(\upsilon + \delta^{(k)}\right)\right\|_2^2\right\}. \tag{2.2}$$

为了使(2.2)式中的系数满足稀疏假设, 在(2.2)式中引入正则项, 在文中考虑使用 LASSO 惩罚函数, 则(2.2)式变为如下(2.3)式：

$$\left(\hat{\upsilon}, \hat{\delta}^{(1)}, \cdots \hat{\delta}^{(K)}\right) \in \underset{\upsilon, \delta^{(1)}, \cdots \delta^{(K)}}{argmin} \begin{Bmatrix} \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\rho_\tau\left(y_i^{(k)} - x_i^{(k)}\left(\upsilon + \delta^{(k)}\right)\right) + \\ \lambda_1\|\upsilon\|_1 \sum_{k=1}^{K}\lambda_2^{(k)}\left\|\delta^{(k)}\right\|_1 \end{Bmatrix}. \tag{2.3}$$

为了方便对(2.2)式进行估计, 对其进行如下变换, 令

$$\theta^{(k)} = \frac{\lambda_2^{(k)}}{\lambda_1}.$$

$$X^*_{n\times(K+1)p} = \begin{bmatrix} x^{(1)} & \frac{x^{(1)}}{\theta^{(1)}} & \cdots & 0_{n_k\times p} \\ \vdots & \vdots & \ddots & \vdots \\ x^{(K)} & 0_{n_k\times p} & \cdots & \frac{x^{(K)}}{\theta^{(k)}} \end{bmatrix}.$$

$$\beta = (\upsilon^T, \left(\theta^{(1)}\delta^{(1)}\right)^T, \cdots, \left(\theta^{(k)}\delta^{(K)}\right)^T)^T.$$

则(2.3)式变为

$$\hat{\beta} = argmin\left\{\frac{1}{n}\|Y - X^*\beta\|_2^2 + \lambda_1\|\beta\|_1\right\}. \tag{2.4}$$

(2.4)式为分层线性回归模型的参数估计式, 分层分位数回归模型的参数估计式将会在第 3 章与第 4 章中给出.

3 响应变量缺失下异方差分层分位数回归模型的估计

3.1 符号

记总体样本容量为n, 样本分为K层, $K \geq 1$. 令n_k表示每层的样本容量, 满足$\sum_{k=1}^{K} n_k = n$. 令$Y = \left(\left(y^{(1)}\right)^T, \cdots, \left(y^{(K)}\right)^T\right)^T$, $y^{(k)}$表示第k层响应变量, $x^{(k)}$表示第k层协变量, 该符号与第 2 章预备知识 2.2 节的符号类似, 与第 4 章中的符号一致, 所以第 3 章之后, 不会在第 4 章中出现单独的符号章节.

3.2 模型的估计方法

3.2.1 分层分位数回归模型

假定$\left\{\left(y_i^{(k)}, x_i^{(k)}\right), i = 1, \cdots, n_k\right\}$为第$k$层独立同分布的样本, 其中$y_i^{(k)}$为响应变量, $x_i^{(k)} = (x_{i1}^{(k)}, \cdots, x_{ip}^{(k)})$为协变量. 给定任意分位数水平$\tau \in (0,1)$, 在给定$x_i^{(k)}$情况下$y_i^{(k)}$的$\tau$分位数定义为

$$Q_{y_i^{(k)}|x_i^{(k)}}(\tau) = \inf\{t: F(t|x_i^{(k)}) \geq \tau\},$$

其中$F(\cdot\,|x_i^{(k)})$为$y_i^{(k)}$的条件分布函数. 则第k层的分位数模型可表示如下, $\beta^{(k)}$表示第k层的回归系数

$$Q_{y_i^{(k)}|x_i^{(k)}}(\tau) = x_i^{(k)}\beta^{(k)}. \tag{3.1}$$

令$\epsilon_i^{(k)} = y_i^{(k)} - Q_{y_i^{(k)}|x_i^{(k)}}(\tau)$, 且$\epsilon_i^{(k)}$满足$p\left(\epsilon_i^{(k)} \leq 0 \middle| x_i^{(k)}\right) = \tau$.

为了表现出各层之间的结构, 对(3.1)式中的回归系数$\beta^{(k)}$进行如下分解：

$$\beta^{(k)} = \upsilon + \delta^{(k)}, \tag{3.2}$$

其中$\upsilon = (\upsilon_1, \cdots, \upsilon_p)$表示不对样本进行分层时的共性回归系数, $\delta^{(k)} = (\delta_1^{(k)}, \cdots, \delta_p^{(k)})$表示第$k$层的个性回归系数, 认为自变量和因变量在不考虑分层信息时具有基础的相关关系υ. 当考虑分层结构时, 第k层的相关关系会υ的基础上变化$\delta^{(k)}$个单位.

该分解方法的核心思想与分层线性回归模型的分解核心思想一致为：要求原始数据在分层前属于一个大类，在分层后各层属于这个大类中的子类，分层后同时包含大类的特征和子类的信息. 因此分层后的回归系数由共性与个性回归系数组合而来.

对于基于 (3.2) 式的$(\upsilon, \delta^{(1)}, \cdots \delta^{(K)})$可以通过如下目标函数进行估计：

$$\left(\hat{\upsilon}, \hat{\delta}^{(1)}, \cdots \hat{\delta}^{(K)}\right) \in \underset{\upsilon, \delta^{(1)}, \cdots \delta^{(K)}}{argmin} \left\{\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\rho_\tau\left(y_i^{(k)} - x_i^{(k)}\left(\upsilon + \delta^{(k)}\right)\right)\right\}. \quad (3.3)$$

其中$\rho_\tau(u) = u(\tau - I(u < 0))$. 为了方便对(3.3)式进行估计，令

$$X_{n\times(K+1)p} = \begin{bmatrix} x^{(1)} & x^{(1)} & \cdots & 0_{n_K\times p} \\ \vdots & \vdots & \ddots & \vdots \\ x^{(K)} & 0_{n_1\times p} & \cdots & x^{(K)} \end{bmatrix}, \quad (3.4)$$

$$\beta = \left(\upsilon^T, \left(\delta^{(1)}\right)^T, \cdots, \left(\delta^{(K)}\right)^T\right)^T. \quad (3.5)$$

则(3.3)式变为

$$\hat{\beta} = argmin\left\{\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(Y_i - X_i\beta)\right\}. \quad (3.6)$$

3.2.2 缺失数据下的分层分位数回归

假定每层的协变量$x^{(k)}$可以被完全观测，而响应变量$y^{(k)}$会出现随机缺失(MAR)的情况，即$y^{(k)}$的缺失与自身无关，只与$x^{(k)}$相关. 当指示变量$\widetilde{\Delta}_i = 1$时，表示$y_i^{(k)}$可以被观测到，当指示变量$\widetilde{\Delta}_i = 0$时，表示$y_i^{(k)}$出现缺失，即

$$p\left(\widetilde{\Delta}_i = 1 \middle| y_i^{(k)}, x_i^{(k)}\right) = p\left(\widetilde{\Delta}_i = 1 \middle| x_i^{(k)}\right) = \tilde{\pi}(x_i^{(k)}),$$

其中$\tilde{\pi}(x_i^{(k)})$称为选择概率函数. 为了方便表示，记$\tilde{\pi}(x_i^{(k)})$为$\tilde{\pi}_i$.

在实际情况中$\tilde{\pi}_i$一般是未知的，通常使用非参数核光滑方法进行估计，但是在高维数据的情况下，非参数核光滑方法的估计精度会随着$x_i^{(k)}$维度的增加而下降，所以本文假定$\tilde{\pi}_i$满足 logistic 模型

$$\tilde{\pi}_i = \frac{exp\,(\alpha_1 + \alpha_2^T x_i^{(k)})}{1 + exp\,(\alpha_1 + \alpha_2^T x_i^{(k)})}.$$

其中$\alpha = (\alpha_1, \alpha_2^T)^T$为未知参数.

在随机缺失(MAR)的假设下，使用逆概率加权(IPW)方法，通过$\frac{\widetilde{\Delta}_i}{\tilde{\pi}_i}$估计第$i$点的权重来减小潜在的偏差，则第$k$层缺失数据下分位数回归的估计参数可由下式求得：

$$\left(\hat{\upsilon},\hat{\delta}^{(k)}\right)\in \underset{\upsilon,\delta^{(k)}}{argmin}\left\{\sum_{i=1}^{n_k}\frac{\widetilde{\Delta}_i}{\tilde{\pi}_i}\rho_\tau\left(y_i^{(k)}-x_i^{(k)}\left(\upsilon+\delta^{(k)}\right)\right)\right\}. \tag{3.7}$$

在本文中，我们假设所有的K个层均拥有相同的选择概率函数$\tilde{\pi}_i$. 下面考虑基于(3.6)式的响应变量随机缺失的分层分位数回归，令

$$\pi_i=\frac{exp\,(\alpha_1+\alpha_3^T X_i)}{1+exp\,(\alpha_1+\alpha_3^T X_i)},$$

$$\Delta_i=\begin{cases}1, Y_i\text{被观测}\\ 0, Y_i\text{缺失}\end{cases}.$$

其中$\alpha_3=(\alpha_2^T,\mathbf{0})^T$，$\mathbf{0}$为$K\times p$维的向量. 则式(3.7)可以变成

$$\hat{\beta}=argmin\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\pi_i}\rho_\tau(Y_i-X_i\beta)\right\}. \tag{3.8}$$

从$\widetilde{\Delta}_i$到Δ_i、$\tilde{\pi}_i$到π_i可以视为是从分层到整体的一个自然的延申，$\tilde{\pi}_i$到π_i的变换中并不会改变选择概率函数$\tilde{\pi}_i$的机制，即$\tilde{\pi}_i$与π_i拥有相同的机制. 响应变量无论是基于(3.7)式分层缺失还是基于(3.8)式的总体缺失，并没有区别，(3.7)式的分层缺失通过(3.4)、(3.5)式的变换后就是(3.8)式的总体缺失.

为了后续待估参数渐近性的证明，先给出证明所需的 8 个条件与 2 个引理.

(1) 随机误差项具有连续可微的密度函数$f_i(\cdot\,|x_i)$，其导数$f_i'(\cdot\,|x_i)$存在于 0 附近的开集中且一致有界. 同时$\max\limits_{1\le i\le n}E\,(\epsilon_i^4)<\infty$.

(2) 存在一个紧集$\mathcal{G}$使得对于任意i，$x_i\in\mathcal{G}\in\mathbb{R}^p$.

(3) 存在$\alpha>0$，使得任意$\pi_i>\alpha$.

(4) Λ、Σ为正定矩阵，易得Σ_1、Λ_1也为正定矩阵，Λ、Σ、Σ_1、Λ_1的定义在定理 3.1 与推论 1 的证明中.

(5) 当$n\to\infty$时，$\lambda/\sqrt{n}\to 0$，$\lambda/n^{1/2-\gamma/2}\to\infty$.

(6) 矩阵 $E\left(X_i'X_i\frac{1}{\pi_i}\phi_\tau(\epsilon_i)^2\right)$为正定矩阵，$\phi_\tau(\epsilon_i)$的定义在定理 3.2 的证明中. 当$n\to\infty$时，$\frac{\max\limits_{1\le i\le n}(X_i)^T X_i}{n}\to 0$.

(7) 随机误差项是相互独立的，$F_i(t)$为其分布函数，假定$F_i(t)$为接近 0 的局

部线性函数，同时满足$F_i(t)=\tau$.

(8) 对于任意μ，$\frac{1}{n}\sum\mathcal{H}_{ni}(\mu^T X_i^*)\to\mathcal{T}(\mu)$，$\mathcal{T}(\cdot)$是一个严格凸函数，且$\mathcal{T}(\cdot)\in[0,\infty)$,$\mathcal{H}_i(\cdot)$的定义在定理 3.2 的证明中.

引理 1 在$x\neq 0$时，$\rho_\tau(x-y)-\rho_\tau(x)=-y\Psi_\tau(x)+\int_0^y[I(x\le t)-I(x\le 0)]dt$成立，其中$\Psi_\tau(x)=\big(\tau-I(x<0)\big)$.

引理 1　即为 Knight 等式，其证明可以参考 Knight(1997).

引理 2 设V是一个正定矩阵，U是一个随机变量，$A_n(s)$是一个凸函数且$A_n(s)=\frac{1}{2}s'Vs+U's+o_p(1)$，则$A_n(s)$的解$\alpha_n\xrightarrow{d}-V^{-1}U$.

引理 2 的证明可以参考 Hjort 等(2011).

定理 3.1　在条件(1)-(4)满足的情况下，有

$$\sqrt{n}(\hat{\beta}-\beta)\xrightarrow{d}N(0,\Lambda^{-1}\Sigma\Lambda^{-1}).$$

又因为$A\hat{\beta}=\hat{\beta}^{(k)}$，则可得

$$\sqrt{n}(\hat{\beta}^{(k)}-\beta^{(k)})\xrightarrow{d}N(0,A\Lambda^{-1}\Sigma\Lambda^{-1}A).$$

其中$A=(E,B)_{p\times(K+1)p}$，$B_{i,j}=\begin{cases}1,\delta_j^{(k)}\neq 0\\ 0,else\end{cases}$，其中$i=k\times j$.

证明：令$\sqrt{n}\big(\hat{\beta}-\beta\big)=\mu$，则有$\beta=\hat{\beta}-\mu/\sqrt{n}$.　记$L_n(\mu)=\sum_{i=1}^n\frac{\Delta_i}{\pi_i}\big[\rho_\tau\big(\epsilon_i-\mu X_i/\sqrt{n}\big)-\rho_\tau(\epsilon_i)\big]$.

由引理 1 可得

$$\begin{aligned}L_n(\mu)&=\sum_{i=1}^n\frac{\Delta_i}{\pi_i}\big[-\mu X_i/\sqrt{n}\,\Psi_\tau(\epsilon_i)\big]+\sum_{i=1}^n\frac{\Delta_i}{\pi_i}\left(\int_0^{\mu X_i/\sqrt{n}}[I(\epsilon_i\le t)-I(\epsilon_i\le 0)]dt\right)\\&\equiv Z1+Z2.\end{aligned}$$

首先研究$Z1$的渐近性质.

$$\begin{aligned}E(Z1)&=E\left(\sum_{i=1}^n\frac{\Delta_i}{\pi_i}\big[-\mu X_i/\sqrt{n}\,\Psi_\tau(\epsilon_i)\big]\right)\\&=-\frac{1}{\sqrt{n}}\sum_{i=1}^n E\left(\frac{\Delta_i}{\pi_i}\mu X_i\Psi_\tau(\epsilon_i)\right)\end{aligned}$$

$$= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} E\left(\mu X_i E(\Psi_\tau(\epsilon_i)|X_i)\right)$$

$$= 0.$$

接下来计算$Z1$的方差，记$Z1i$=$\frac{\Delta_i}{\pi_i}\mu X_i\Psi_\tau(\epsilon_i)$，　则有

$$Var(Z1i) = Var\left(\frac{\Delta_i}{\pi_i}\mu X_i\Psi_\tau(\epsilon_i)\right)$$

$$= \mu' E\left(X_i' X_i \frac{1}{\pi_i}\Psi_\tau(\epsilon_i)^2\right)\mu$$

$$= \mu'\Sigma\mu.$$

通过中心极限定理可得$Z1 \xrightarrow{d} N(0, \mu'\Sigma\mu)$.

接下来研究$Z2$的渐近性质，记$Z2i = \int_0^{\mu X_i/\sqrt{n}}[I(\epsilon_i \le t) - I(\epsilon_i \le 0)]dt$，则有

$$E\left(\sum_{i=1}^{n}\frac{\Delta_i}{\pi_i}Z2i\right) = \sum_{i=1}^{n} E\left(\int_0^{\mu x_i/\sqrt{n}}[I(\epsilon_i \le t) - I(\epsilon_i \le 0)]dt\right)$$

$$= \sum_{i=1}^{n} E\left(\int_0^{\mu X_i/\sqrt{n}}\left(F_i(t) - F_i(0)\right)dt\right)$$

$$= \sum_{i=1}^{n} E\left(\int_0^{\mu X_i/\sqrt{n}}\left(f_i(0)t - f_i'(t^*)\frac{t^2}{2}\right)dt\right).$$

其中t^*在 0 到t之间.

由条件 1 可得

$$\sum_{i=1}^{n} E\left(\int_0^{\mu X_i/\sqrt{n}} f_i(0)t dt\right) = \frac{1}{2}\mu'\frac{1}{n}\sum_{i=1}^{n} E(f_i(0)X_i'X_i)\mu = \frac{1}{2}\mu'\Lambda\mu.$$

由条件 1 与条件 2 可得

$$\sum_{i=1}^{n} E\left(\int_0^{\mu X_i/\sqrt{n}} f_i'(t^*)\frac{t^2}{2}dt\right) \le \frac{C}{3\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^{n}\|\mu\|^3\right) \to 0.$$

有条件 2 与条件 3 可得

$$Var\left(\sum_{i=1}^{n}\frac{\Delta_i}{\pi_i}Z2i\right) \le \sum_{i=1}^{n} E\left(\frac{\Delta_i}{{\pi_i}^2}\left(\int_0^{\mu X_i/\sqrt{n}}[I(\epsilon_i \le t) - I(\epsilon_i \le 0)]dt\right)^2\right)$$

$$\le C\sum_{i=1}^{n} E\left(\left(\int_0^{\mu X_i/\sqrt{n}}[I(\epsilon_i \le t) - I(\epsilon_i \le 0)]dt\right)^2\right)$$

$$\leq \frac{C\|\mu\|}{\sqrt{n}}\sum_{i=1}^{n} E\left(\int_0^{\mu X_i/\sqrt{n}} [I(\epsilon_i \leq t) - I(\epsilon_i \leq 0)]dt\right) \to 0.$$

则可得$\sum_{i=1}^{n}\frac{\Delta_i}{\pi_i} Z2i \xrightarrow{p} \frac{1}{2}\mu'\Lambda\mu.$

综上所述, 由引理 2 与 Cramér-Wold 定理可得$\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} N(0, \Lambda^{-1}\Sigma\Lambda^{-1})$.

3.2.3 异方差假设下缺失数据的分层分位数回归

当每层误差项$\epsilon_i^{(k)}$不满足独立同分布的假设时, 将独立同分布的误差项条件放宽, 考虑异方差条件下的缺失数据的分层分位数回归系数的渐近性.

定理 3.2 在条件(2)-(4)、(6)-(8)满足的情况下,

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \xrightarrow{d} -\mu W + \mathcal{T}(\mu).$$

且$-\mu W + \mathcal{T}(\mu)$拥有唯一解, 其中$W$服从均值为 0 的多元正态分布.

证明： 定义$\mathcal{H}_{ni}(t) = \int_0^t \sqrt{n}\,(F_i\left(\frac{s}{\sqrt{n}}\right) - F_i(0))ds.$

首先对$\rho_\tau(\cdot)$进行改写如下,

$$\rho_\tau(r) = \frac{|r|}{2} + (\tau - \tfrac{1}{2})r.$$

$$\begin{aligned} L_n(\mu) &= \sum_{i=1}^{n}\frac{\Delta_i}{\pi_i}\left[\rho_\tau\left(\epsilon_i - \mu X_i^*/\sqrt{n}\right) - \rho_\tau(\epsilon_i)\right] \\ &= \sum_{i=1}^{n}\frac{\Delta_i}{\pi_i}\frac{-X_i^*\mu}{\sqrt{n}}\left(\frac{sign(\epsilon_i)}{2} + \left(\tau - \frac{1}{2}\right)\right) \\ &\quad + \sum_{i=1}^{n}\frac{\Delta_i}{\pi_i}\left(\int_0^{\mu X_i^*/\sqrt{n}} [I(\epsilon_i \leq t) - I(\epsilon_i \leq 0)]dt\right) \\ &\equiv L_{1n}(\mu) + L_{2n}(\mu). \end{aligned}$$

由中心极限定理与条件(6)可得, $L_{1n}(\mu) \xrightarrow{d} -\mu W$且$W \sim N(0, C_1), C_1 = E\left(X_i' X_i \frac{1}{\pi_i}\phi_\tau(\epsilon_i)^2\right)$, $\phi_\tau(\epsilon_i) = \frac{sign(\epsilon_i)}{2} + \left(\tau - \frac{1}{2}\right)$.

下面我们证明$L_{2n}(\mu)$.

$$L_{2n}(\mu) = \sum_{i=1}^{n} L_{2ni}(\mu).$$

令$v_i = \mu X_i^*$对$L_{2n}(\mu)$求期望得

$$
\begin{aligned}
E\left(L_{2ni}(\mu)\right) &= \sum_{i=1}^{n} E\left(L_{2ni}(\mu)\right) \\
&= \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \int_0^{v_i} \left(F_i\left(\frac{t}{\sqrt{n}}\right) - F_i(0) \right) dt \\
&= \frac{1}{n} \sum_{i=1}^{n} \sqrt{n} \int_0^{v_i} \left(F_i\left(\frac{t}{\sqrt{n}}\right) - F_i(0) \right) dt \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathcal{H}_{ni}(\mu) \\
&\to \mathcal{T}(\mu).
\end{aligned}
$$

最后一个等号由条件(8)可得.

接下来求$L_{2n}(\mu)$的方差, 结合$\mathcal{T}(\mu) < \infty$与定理 3.1 证明可得$Var\left(L_{2n}(\mu)\right) \to 0$.

综上可得$L_n(\mu) \to L(\mu) = -\mu W + \mathcal{T}(\mu)$. 由于$\mathcal{T}(\mu)$是严格的凸函数, $L(\mu)$拥有唯一的解, 则定理 3.2 得证.

推论 3.1 通过定理 3.1 与定理 3.2 可得, 在条件(2)-(4)、(6)-(8)满足的情况下定理 3.1 依旧成立.

证明：结合定理 3.1 的证明过程易得推论 3.1 成立.

3.2.4 缺失数据下的分层分位数回归的变量选择

如果每层数据的维度p较大, 则(3.8)式可能无法进行精确的估计, 导致误差的产生. 为了更精确的对(3.8)式进行估计, 本节引入变量选择的方法. 首先考虑 LASSO 惩罚函数, 构建如下目标函数

$$
\left(\hat{\upsilon}, \hat{\delta}^{(k)}\right) \in \underset{\upsilon, \delta^{(k)}}{argmin} \left\{ \begin{matrix} \sum_{i=1}^{n_k} \frac{\dot{\lambda}_i}{\dot{\pi}_i} \rho_\tau \left(y_i^{(k)} - x_i^{(k)} \left(\upsilon + \delta^{(k)} \right) \right) + \lambda_1 \|\upsilon\|_1 + \\ \sum_{k=1}^{K} \lambda_2^{(k)} \left\| \delta^{(k)} \right\|_1 \end{matrix} \right\}. \tag{3.9}
$$

由于采用了ℓ_1范数对υ和$\delta^{(k)}$进行了惩罚, 因此这两个参数共同决定了$\beta^{(k)}$的稀疏性. 并且λ_1越大, $\|\upsilon\|_1$就会越小, 当λ_1大到一定程度时, $\|\upsilon\|_1$会趋向于 0, 此时$\beta^{(k)} = \delta^{(k)}$, 即共性回归系数为 0, 此时就变成了 K 个相互独立的层数的 LASSO 估计. 同理, 当$\lambda_2^{(k)}$大到一定程度时, $\left\|\delta^{(k)}\right\|_1$会趋向于 0, 此时$\beta^{(k)} = \upsilon$,即个性回归系数为 0, 此时就变成了不考虑分层而对总体直接进行的 LASSO 估计.

因此上述两种情况均是本文方法的特殊情况.

为了构造方便估计的诸如(3.8)式那样便于估计的式子, 我们采取矩阵变换的方法. 首先对(3.9)式做如下变换, 令$\theta^{(k)}=\frac{\lambda_2^{(k)}}{\lambda_1}$则:

$$\lambda_1\|\upsilon\|_1+\sum_{k=1}^{K}\lambda_2^{(k)}\left\|\delta^{(k)}\right\|_1=\lambda_1\|\beta^*\|_1.$$

$$X^*_{n\times(K+1)p}=\begin{bmatrix} x^{(1)} & \frac{x^{(1)}}{\theta^{(1)}} & \cdots & 0_{n_k\times p} \\ \vdots & \vdots & \ddots & \vdots \\ x^{(K)} & 0_{n_k\times p} & \cdots & \frac{x^{(K)}}{\theta^{(K)}} \end{bmatrix}.$$

可得

$$\hat{\beta}^*=argmin\left\{\sum_{i=1}^{n}\frac{\Delta_i}{\pi_i}\rho_\tau(Y_i-X_i^*\beta)+\lambda_1\|\beta^*\|_1\right\}. \tag{3.10}$$

其中$\beta^*=(\upsilon^T,\left(\theta^{(1)}\delta^{(1)}\right)^T,\cdots,\left(\theta^{(K)}\delta^{(K)}\right)^T)$.

推论 3.2 可以选择满足 Oracle 性质的 ADALASSO 惩罚函数来改进式(3.10). 在条件(1)-(5)满足的情况下, 有

(1) $p(\hat{\beta}_{12}^*=0)\xrightarrow{p}1$.

(2) $\sqrt{n}(\hat{\beta}_{11}^*-\beta_{11}^*)\underset{d}{\rightarrow}N(0,A\Lambda_1^{-1}\Sigma_1\Lambda_1^{-1}A)$.

又因为$A\hat{\beta}_{11}^*=\hat{\beta}_{11}^{(k)}$, 则可得

$$\sqrt{n}(\hat{\beta}_{11}^{(k)}-\beta_{11}^{(k)})\xrightarrow{d}N(0,A\Lambda_1^{-1}\Sigma_1\Lambda_1^{-1}A).$$

证明:

$$\left(\hat{v},\hat{\delta}^{(k)}\right)\in\underset{v,\delta^{(k)}}{argmin}\left\{\sum_{i=1}^{n_k}\frac{\widetilde{\Delta}_i}{\tilde{\pi}_i}\rho_\tau\left(y_i^{(k)}-x_i^{(k)}\left(\upsilon+\delta^{(k)}\right)\right)+\lambda_1\omega_1|v|+\sum_{k=1}^{K}\lambda_2^{(k)}\,\omega_2^{(k)}\left|\delta^{(k)}\right|\right\}.$$

使用同式(3.9)的转换方法可得

$$\hat{\beta}^*=argmin\left\{\sum_{i=1}^{n}\frac{\Delta_i}{\pi_i}\rho_\tau(Y_i-X_i^*\beta)+\lambda_1\omega^*|\beta^*|\right\}.$$

其中$\omega^*=(\omega_1^T,\left(\omega_2^{(1)}\right)^T,\cdots,\left(\omega_2^{(K)}\right)^T)^T=\left|\hat{\beta}^*\right|^{-\gamma}$, $\omega_1=|\hat{v}|^{-\gamma},\omega_2^{(k)}=\left|\hat{\delta}^{(k)}\right|^{-\gamma}$, 令$L(\beta^*)=\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\pi_i}\rho_\tau(y_i-X_i^*\beta^*)+\sum_{j=1}^{(K+1)p}p\left(\left|\beta_j^*\right|\right)$, 对$L(\beta^*)$求偏导可得

$$\frac{\partial L(\beta^*)}{\partial\beta_j^*}=\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\pi_i}X_{ij}^*\Psi_\tau(\epsilon_i)-\frac{1}{n}\sum_{i\in\mathcal{D}}^{n}\frac{\Delta_i}{\pi_i}X_{ij}^*\Psi_\tau(\epsilon_i)[a_i+(1-\tau)]+$$

$$p'\left(\left|\beta_j^*\right|\right)sgn(\beta_j^*).$$

其中若$y_i - X_i^*\beta^* = 0$, 则$a_i = 0$, 若$\mathcal{D} = \{i: y_i - X_i^*\beta^* \neq 0\}$, 则$a_i \in [\tau - 1, \tau]$.

结合定理 3.1, 当$n \to \infty$时,$\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\pi_i}X_{ij}^*\Psi_\tau(\epsilon_i) \to 0$、$\frac{1}{n}\sum_{i\in\mathcal{D}}^{n}\frac{\Delta_i}{\pi_i}X_{ij}^*\Psi_\tau(\epsilon_i)[a_i + (1-\tau)] \to 0$.

所以$\frac{\partial L(\beta^*)}{\partial \beta_j^*}$的符号由$p'(|\beta_j^*|)sgn(\beta_j^*)$来决定.

对于 ADALASSO 惩罚函数,

$$p'(|\beta_j^*|) > 0.$$

则$p'(|\beta_j^*|)sgn(\beta_j^*)$的符号取决于$\beta_j^*$的符号, 则结合 Bai(2020)中定理 2 的证明可得

$$p(\hat{\beta}_{12}^*) = 0 \xrightarrow{p} 1.$$

下面证明推论 3.2(2), 令

$$V(\mu) = \sum_{i=1}^{n}\frac{\Delta_i}{\pi_i}[\rho_\tau(\epsilon_i - \mu X_i^*/\sqrt{n}) - \rho_\tau(\epsilon_i)] + n\sum_{j=1}^{(p+1)K}[p(|\beta_j^* + \mu_j/\sqrt{n}|) - p(|\beta_j^*|)]$$
$$\equiv V_1(\mu) + V_2(\mu).$$

首先考虑$V_2(\mu)$

如果$\beta_j^* \neq 0$, 则由条件(5)可得

$$|\lambda_1\omega^*(|\beta_j^* + \mu_j/\sqrt{n}| - |\beta_j^*|)| \leq \lambda_1\omega^*|\mu_j/\sqrt{n}| \to 0.$$

如果$\beta_j^* \neq 0$, 则由条件(5)与$\sqrt{n}\hat{\beta}_j^* = O_p(1)$可得,

$$|\lambda_1\omega^*(|\beta_j^* + \mu_j/\sqrt{n}| - |\beta_j^*|)| = \lambda_1\omega^*|\mu_j/\sqrt{n}|$$
$$= \frac{\lambda_1}{n^{1/2-\gamma/2}}\frac{|\mu_j|}{|\sqrt{n}\hat{\beta}_j^*|^\gamma}$$
$$= \begin{cases}\infty, \mu_j \neq 0, \\ 0, \mu_j = 0.\end{cases}$$

综上结合定理 3.1 由 Slutsky 定理可得

$$\sqrt{n}(\hat{\beta}_{11}^* - \beta_{11}^*) \xrightarrow{d} N(0, A\Lambda_1^{-1}\Sigma_1\Lambda_1^{-1}A).$$

3.3 模拟研究

蒙特卡洛模拟参数设置如下：$k = 10, p = 50$, $n_k \in \{40,50,100\}$, 观测矩阵

$X\sim N(0,\Sigma)$, 其中Σ是$p\times p$维的 T 型矩阵, $\Sigma=(0.5^{|i-j|})_{1\le i,j\le p}$, 其中多个$n_k$的值是为了研究$p/n_k$的比值对估计精度的影响. 接下来开始构造$\beta^{(k)}$, 首先从$p$个维度中选择 20 个维度记作集合$P_0$, $P_0\in\{1,2,\cdots,p\}$, 同时令$i\in\{1,2,\cdots,p\}$, 当$i\notin P_0$时, $\beta_i^{(k)}=0$, 当$i\in P_0$时, 再从P_0中抽取 10 个元素, 组成集合P_1, 当$i\in P_0$且$i\notin P_1$时, $\beta^{(k)}$分两种情况构造, 当$k<2$时, $\beta_i^{(k)}=1$, 当$\mathrm{k}>2$时, $\beta_i^{(k)}=1+\gamma^k$; 当$i\in P_1$时, $\beta^{(k)}$也分两种情况构造, 当$k<2$时, $\beta_i^{(k)}=1+\gamma^k$, 当$k>2$时, $\beta_i^{(k)}=1$, 上述过程中令$\gamma^k=0.1\times K^{1/2}$. 通过上述过程, 构造共性回归系数是 1、个性回归系数是 0 或γ^k. 虽然模拟构造麻烦, 但保证了回归系数的稀疏性. 记$\beta^\circ=(\beta^{(1)},\cdots,\beta^{(10)})$, $\hat{\beta}^\circ$为其估计值. 随机误差项$\varepsilon^{(k)}$由如下 4 种方法构造, 其中第 4 种构造方法模拟异方差情况：

(1) $\epsilon^{(k)}\backsim t(3)$.

(2) $\epsilon^{(k)}\backsim N(0,1)$.

(3) $\epsilon^{(k)}\backsim N(0,1)$, 随机选出 40%, 其中的一半取 5, 另一半取-5.

(4) $\epsilon^{(k)}=x^{(k)}\beta^{(k)}\circ\varepsilon$, $\varepsilon\backsim N(0,1)$.

数据缺失通过如下三种概率选择函数进行模拟：

$$logit_1\big(P(\Delta_i=1|X_i^*)\big)=2.65+0.75X_{i1}^*+0.75X_{i2}^*,$$

$$logit_2\big(P(\Delta_i=1|X_i^*)\big)=1.05+0.75X_{i1}^*+0.75X_{i2}^*,$$

$$logit_3\big(P(\Delta_i=1|X_i^*)\big)=-1.05+0.75X_{i1}^*+0.75X_{i2}^*.$$

$logit_1\sim logit_3$分别对应 8%、30%与 67%的缺失率.

为了比较不同方法变量选择和参数估计的准确性, 我们选取 TPR、FPR、ℓ_2这 3 种指标来衡量模型的估计性能, 定义方法如下：

$$\mathrm{TPR}=\frac{\#(i:\hat{\beta}_i^\circ\neq 0\ \text{且}\ \beta_i^\circ\neq 0)}{\#(i:\beta_i^\circ\neq 0)},$$

$$\mathrm{FPR}=\frac{\#(i:\hat{\beta}_i^\circ\neq 0\ \text{且}\ \beta_i^\circ=0)}{\#(i:\beta_i^\circ=0)},$$

$$\ell_2=\left\|\hat{\beta}_i^\circ-\beta_i^\circ\right\|_2.$$

TPR、FPR 分别表示正确估计非 0 回归系数的比例与错误估计非 0 回归系数的比例. TPR 越大表示模型的估计精度越高, 同理 FPR 越小表示模型的估计精度越高. ℓ_2用来衡量模型回归系数的误差, ℓ_2越小表示模型精度越高. 在模拟研究中无论是$\beta^{(k)}$的构造还是模型的拟合, 均不考虑截距项. 为了表明本文提出方法的优劣性, 我们把不含有缺失值的分层分位数回归模型与分层线性回归模型作为对比模型, 分别记为 OMNI 与 LR. 分层分位数回归模型记为 QR.

表 3.1 $p/n_k = 1$时模型估计结果

τ	模型	误差项	缺失率	TPR	FPR	ℓ_2
0.5	OMNI	(1)	-	1.000	0.020	1.942
	QR		8%	1.000	0.023	1.987
	LR		8%	1.000	0.217	2.192
	QR		30%	1.000	0.067	2.304
	LR		30%	1.000	0.237	2.575
	QR		67%	1.000	0.223	4.396
	LR		67%	1.000	0.363	4.220
	OMNI	(2)	-	1.000	0.038	1.694
	QR		8%	1.000	0.063	1.994
	LR		8%	1.000	0.093	1.497
	QR		30%	1.000	0.123	2.314
	LR		30%	1.000	0.240	1.705
	QR		67%	1.000	0.250	3.913
	LR		67%	1.000	0.307	2.474
	OMNI	(3)	-	1.000	0.044	1.873
	QR		8%	1.000	0.080	2.108
	LR		8%	1.000	0.170	3.055
	QR		30%	1.000	0.087	2.441
	LR		30%	1.000	0.237	3.705
	QR		67%	1.000	0.217	5.236
	LR		67%	1.000	0.287	5.979

观察表 3.1 的结果可以看出, 基于 QR、LR 与 OMNI 方法的 TPR 相等, 均为

1. 表明 3 种模型均可以正确估计非 0 回归系数. 在同一误差项下, 基于 OMNI 方法的 FPR 最低, 基于 LR 方法的 FPR 最高. 表明相较于 LR 方法, OMNI 方法可以减少错误估计非 0 回归系数的个数. 并且在同一误差项的相同的缺失率下, 基于 QR 方法的 FPR 低于基于 LR 方法的 FPR, 表明 QR 方法相较于 LR 方法, 也可以减少错误估计非 0 回归系数的个数.

只有在误差项服从正态分布时, 基于 QR、OMNI 方法的估计误差ℓ_2大于 LR. 其余的情况下 QR、OMNI 方法的表现均要优于 LR 方法, 且 OMNI 方法优于 QR 方法. 同时, OMNI、QR 与 LR 方法均在误差项服从正态分布时估计精度最高. 当误差项为(3)时, OMNI 与 QR 方法的优势会更加明显.

随着缺失率的增加, 基于 QR 与 LR 方法的 FPR 与ℓ_2都会呈现出上升的趋势, 尤其是当缺失率为 67%时, FPR 与ℓ_2会迅速增大. 因此随着缺失率的增加, 基于 OMNI、QR 与 LR 方法的估计精度均会降低.

表 3.2 $p/n_k < 1$时模型估计结果

τ	模型	误差项	缺失率	TPR	FPR	ℓ_2
0.5	OMNI	(1)	-	1.000	0.040	1.246
	QR		8%	1.000	0.053	1.488
	LR		8%	1.000	0.103	1.254
	QR		30%	1.000	0.210	1.572
	LR		30%	1.000	0.230	1.748
	QR		67%	1.000	0.250	2.519
	LR		67%	1.000	0.310	3.631
	OMNI	(2)	-	1.000	0.023	1.209
	QR		8%	1.000	0.034	1.617
	LR		8%	1.000	0.087	1.167
	QR		30%	1.000	0.167	1.797
	LR		30%	1.000	0.337	1.219
	QR		67%	1.000	0.130	2.565
	LR		67%	1.000	0.623	2.122
	OMNI	(3)	-	1.000	0.047	1.382
	QR		8%	1.000	0.067	1.547
	LR		8%	1.000	0.333	1.773
	QR		30%	1.000	0.073	1.582
	LR		30%	1.000	0.337	2.396
	QR		67%	1.000	0.106	4.048
	LR		67%	1.000	0.490	5.887

观察表 3.2 可以得出与表 3.1 相同的结论, 同时因为$p/n_k < 1$, 表 3.2 中的ℓ_2与 FPR 相较于表 3.1 来说较小.

表 3.3 $p/n_k > 1$时模型估计结果

τ	模型	误差项	缺失率	TPR	FPR	ℓ_2
0.5	OMNI	(1)	-	1.000	0.073	2.071
	QR		8%	1.000	0.093	2.064
	LR		8%	1.000	0.130	2.807
	QR		30%	1.000	0.167	2.995
	LR		30%	1.000	0.313	3.062
	QR		67%	1.000	0.237	5.674
	LR		67%	1.000	0.410	6.607
	OMNI	(2)	-	1.000	0.077	1.779
	QR		8%	1.000	0.033	2.094
	LR		8%	1.000	0.097	1.582
	QR		30%	1.000	0.223	2.462
	LR		30%	1.000	0.307	1.718
	QR		67%	1.000	0.183	3.557
	LR		67%	1.000	0.467	3.094
	OMNI	(3)	-	1.000	0.070	1.980
	QR		8%	1.000	0.059	1.881
	LR		8%	1.000	0.107	3.250
	QR		30%	1.000	0.083	3.279
	LR		30%	1.000	0.167	4.083
	QR		67%	0.963	0.119	8.258
	LR		67%	0.985	0.373	8.371

观察表 3.3 也可以得出与表 3.1 相同的结论, 同时因为$p/n_k > 1$, 表 3.3 中的ℓ_2与 FPR 相较于表 3.1 来说较大.

综上所述, 表 3.1 得出的结论适用于表 3.2、表 3.3, 即p/n_k的比值不改变 QR 与 LR 方法之间的优劣性, 仅会影响 QR 与 LR 方法的估计误差的大小, 并且当$p/n_k > 1$时, 3 种模型的估计的误差较大, $p/n_k < 1$时, 估计误差较小. 当模型响应变量缺失且误差项不服从正态分布时, QR 方法相较于 LR 方法具有一定的优越性.

接下来, 考虑异方差情况下, QR、OMNI 与 LR 三种方法的估计性能.

表 3.4 异方差情况下模型估计结果

p/n_k	τ	模型	缺失率	TPR	FPR	ℓ_2
<1	0.5	OMNI	-	1.000	0.057	5.199
		QR	8%	0.992	0.093	5.652
		LR	8%	0.975	0.146	5.760
		QR	30%	0.984	0.107	7.865
		LR	30%	0.910	0.270	7.989
		QR	67%	0.730	0.133	11.486
		LR	67%	0.485	0.687	12.145
$=1$	0.5	OMNI	-	1.000	0.050	6.785
		QR	8%	0.980	0.090	8.584
		LR	8%	0.950	0.200	9.192
		QR	30%	0.905	0.180	9.294
		LR	30%	0.800	0.203	9.933
		QR	67%	0.645	0.313	13.039
		LR	67%	0.650	0.620	17.749
>1	0.5	OMNI	-	1.000	0.063	8.707
		QR	8%	0.976	0.095	8.913
		LR	8%	0.943	0.230	9.575
		QR	30%	0.810	0.113	9.926
		LR	30%	0.780	0.318	10.625
		QR	67%	0.408	0.081	14.950
		LR	67%	0.368	0.114	14.587

观察表 3.4 可以发现, 在异方差情况下, 同一p/n_k的水平下, 随着缺失率的增加, 基于 OMNI、QR 与 LR 方法的 FPR 与ℓ_2都会增大, 基于 QR 与 LR 方法的 TPR 会减小. 并且缺失率对估计精度的影响相较于表 3.1~表 3.3 有极大的提高, 通过对比 OMNI、QR 与 LR 方法可以发现 TPR 对缺失率的变化最为明显, 随着缺失率的增加基于 QR 与 LR 方法的 TPR 会迅速下降.

在异方差情况下p/n_k的比值对估计的影响也较大. 同一模型下, 随着p/n_k的增加, 基于 FPR 方法与ℓ_2会增大、TPR 会减小. 当$p/n_k > 1$时, 这种变化幅度随着缺失率的增加更加的明显, 当缺失率为 67%时, 基于 QR 方法的ℓ_2甚至会略大于基于 LR 方法的ℓ_2. 同时也发现随着缺失率增加 FPR 会出现减少的现象. 对于这种现象的分析, 会在表 3.5 的分析中给出.

表 3.5 为$p/n_k > 1$时, 异方差情况下基于各方法的非 0 系数个数表. 表中的元素为非 0 系数个数的平均数. 括号内为非 0 系数的极差, 可以反映模型的稳健性. 模型中真实非 0 系数为 200.

表 3.5 异方差情况下各模型非 0 系数个数

缺失率	QR	LR
8%	190.4 (63)	260.8 (71)
30%	164.8 (137)	251.3 (154)
67%	116.5 (142)	183.2 (325)

从表 3.5 中可以发现, 随着缺失率的增加, 基于 QR 方法的非 0 系数个数逐渐减少, 并且越来越偏离 200. 而基于 LR 方法的非 0 系数个数同样逐渐减少, 但是越来越接近 200. 因此在所有缺失率下, QR 方法倾向于估计较少的非 0 系数, LR 方法倾向于估计较多的非 0 系数, 所以基于 QR 方法的 FPR 会小于 LR. 随着缺失率的增加, 两种模型估计非 0 系数个数的极差均增加, 导致模型估计精度下降. 同时非 0 系数个数急剧的减小, 导致了 FPR 出现下降的情况. 上述情况在基于 LR 方法上的表现较为明显. 由于基于 QR 方法的非 0 系数在缺失率为 67%时减小过大, 这可能导致了基于 QR 方法的ℓ_2略大于 LR.

3.4 实例分析

在这一节中, 将使用文中所提出的方法分析 THP-1 人骨髓单核细胞白血病细胞中分化为巨噬细胞数据. 其中这些因子的表达已经通过分析每个时间点的 120 个不同的单细胞来测量. 该数据来自于 Kouno 等(2013)发表的论文, 其中包含 45 种转录因子在 8 种不同时间点(H0、H1、 H6、H12、H24、H48、H72 与 H96)的表达水平. 可以将整个时间段看做一个大类, 截取的 8 个时间点看作大类下的子类. 因此符合本文对于回归系数分解的要求. 故而, 可以通过不同的时间段将数据分为 8 层.

表 3 .6 THP-1 人骨髓单核细胞白血病细胞中分化为巨噬细胞数据正态性分析表

转录因子	CBFB	CEBPB	CEBPD	EGR2	ELK1	ETS1	FLI1	FOS	FOSB	HOXA10	HOXA13
偏度	-0.728	0.560	1.450	1.342	0.663	0.191	-0.528	0.585	2.692	0.073	0.081
峰度	5.684	5.977	14.678	8.775	1.823	2.949	11.015	5.162	11.358	3.868	4.651
转录因子	IRF8	JUN	KLF10	KLF13	LMO2	MAFB	MYB	MYEF2	NFATC1	NFATC2	NFE2L1
偏度	-0.140	1.106	2.781	2.630	1.345	1.226	0.066	-0.179	0.812	-0.456	-0.032

续表 3.6

峰度	1.644	11.803	13.634	27.821	7.828	6.644	2.664	2.179	2.819	3.445	3.124
转录因子	NFYA	NFYC	PPARD	PPARG	PRDM1	RREB1	RUNX1	RXRB	SMAD3	SMAD4	SNAI1
偏度	3.184	0.306	0.876	0.926	0.598	0.082	0.401	-0.738	-0.495	-0.987	0.458
峰度	43.539	6.909	5.316	4.892	7.255	6.687	8.285	5.482	2.308	9.350	5.152
转录因子	SNAI3	SP3	SPI1	SPIB	STAT1	TCF3	TCFL5	TFPT	TRIM28	UHRF1	VDR
偏度	1.248	0.701	2.234	1.283	1.049	-0.811	-0.666	-0.678	0.032	0.198	1.023
峰度	8.833	5.435	16.397	4.124	6.816	4.663	2.443	3.600	3.048	2.838	7.089

观察表 3.6 可以发现, 30%的转录因子的偏度大于 1, 仅有 9%的转录因子的偏度小于 0.1. 70%的转录因子的偏度为正数, 75%的转录因子的峰度大于 3 且 NFYA、KLF13 等转录因子的偏度为 43.539、27.821 远大于 3. 上述结果表明数据集是非对称分布的且呈现出右偏的特性, 同时大部分转录因子的分布也呈现出尖峰的特性.

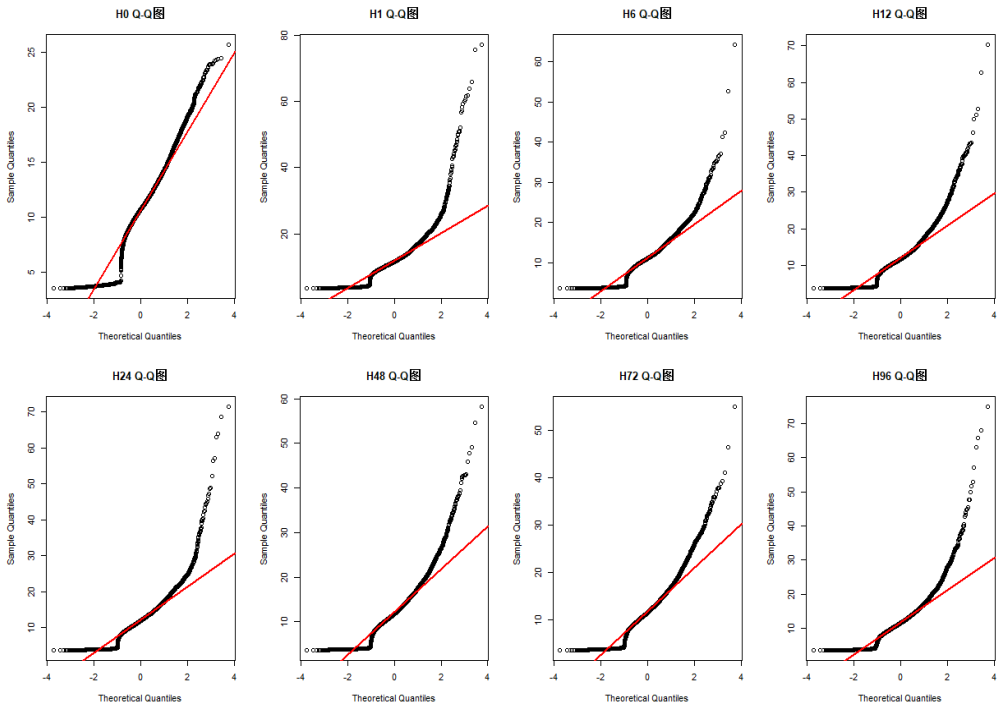

图 3.1 不同时段下的正态性检验 Q-Q 图

通过观察图 3.1 不同时间段数据的 Q-Q 图发现, 8 个时间点的数据均不服从

正态分布. 结合表 3.6 与图 3.1 可以发现该数据集不服从正态分布, 其分布满足尖峰厚尾的特征, 可以使用基于分位数回归模型的 QR 模型进行估计. 因此数据集符合本文提出的在异方差情况下的分层分位数回归的前提假设. 本节研究 BCL6 转录因子与其余 44 个转录因子之间的表达关系. 通过 $logit_4\left(P(\Delta_i = 1|X_i^*)\right) = 0.65 + 0.09X_{i1}^* + 0.05X_{i2}^*$来模拟 BCL6 转录因子的缺失, $logit_4$的缺失率为 10%.

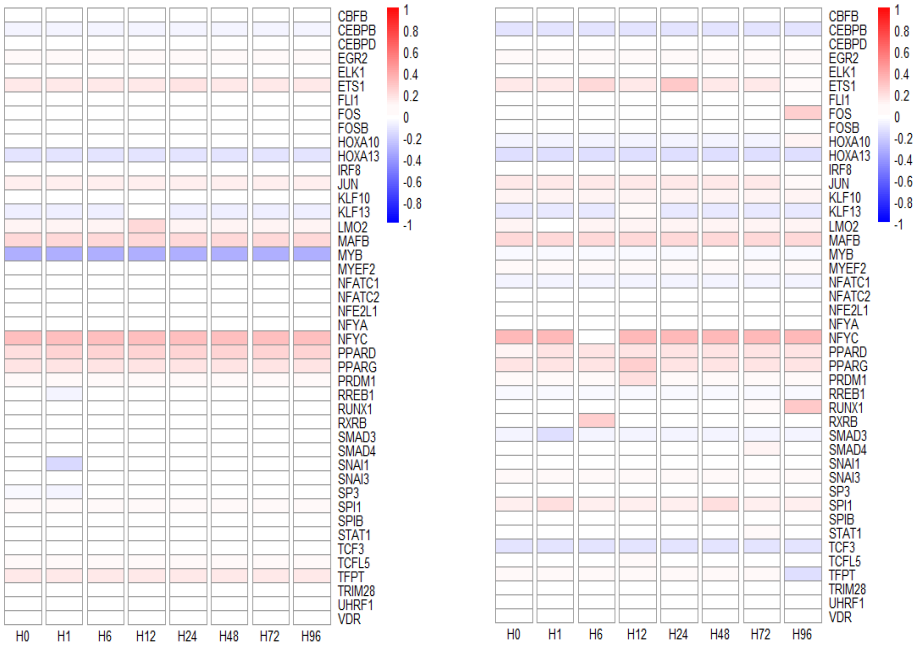

图 3.2 不同模型转录因子表达图(左边为 QR 模型, 右边为 LR 模型)

由图 3.2 的结果可以得出, 在分层分位数回归模型下, BCL6 转录因子仅与 MYB 转录因子呈现出较强的负相关关系, 与 CEBPB、HOXA13、KLF13 转录因子呈现出较弱的负相关关系. 在分层线性模型下, BCL6 转录因子与 MYB、HOXA10、SMAD3、NFATC1、REREB1 转录因子呈现出较弱的负相关关系, 与 TCF3、HOXA13、KLF13、CEBPB 呈现出较强的负相关关系. 2 种分层模型选择出与 BCL6 转录因子有正相关关系的转录因子大致相同为 TFPT、SPL1、NFYC、PPARD、NFYC、MAFB、LMO2、JUN、ETS1、EGR2 这 10 个转录因子.

根据 Kouno 等(2013)的研究以及查阅资料发现, BCL6 转录因子仅与 MYB 转录因子呈现出明显的负相关关系, 与 MAFB、PPARG、SPL1、EGR2、KLF10 转

录因子呈现出明显的正相关关系. 结合实例分析的结果可以发现分层分位数回归模型明显优于分层线性回归模型. 同时也可以发现, 分层线性回归模型会选择出诸如 RXRB(H6)、RUNX1(H96)这样在某一时段的转录因子, 而分层分位数回归模型倾向于减少这样的干扰. 综上所述, 分层分位数回归模型在该数据中表现更好.

4 卷积平滑分位数回归在分层数据中的应用

本章使用的符号与模型的变换过程类似于第 3 章, 所以相似的过程在这里就不在重复.

4.1 分层卷积平滑分位数回归模型

为了提升模型的估计精度与估计速度, 本文使用 He 等(2021)、Tan 等(2022)中提出的卷积平滑方法来改进第 3 章中的(3.3)式. 为了与第 3 章中的符号进行区分, 将第 3 章(3.3)式中的$\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i^{(k)}-x_i^{(k)}\underbrace{\left(\upsilon+\delta^{(k)}\right)}_{\beta^{(k)}}\right)$记为$\hat{Q}(\beta^{(k)})$, 令$\left(\beta^{(k)}\right)^*=\upsilon^*+\delta^{(k)^*}$, $y_i^{(k)}=x_i^{(k)}(\beta^k)^*+\epsilon_i^{(k)}$.

$F_{\epsilon_i^{(k)}|x_i^{(k)}}(\cdot)$是给定$x_i^{(k)}$关于$\epsilon_i^{(k)}$的条件分布函数, 则损失函数$Q\left(\beta^{(k)}\right)=E_{x_i^{(k)}}\{\int_{-\infty}^{\infty}\rho_\tau\left(\mu-\langle x_i^{(k)},\beta^{(k)}-\left(\beta^{(k)}\right)^*\rangle\right)dF_{\epsilon_i^{(k)}|x_i^{(k)}}(\mu)\}$. 假设$F_{\epsilon_i^{(k)}|x_i^{(k)}}(\cdot)$是充分平滑的且$Q\left(\beta^{(k)}\right)$是在$\left(\beta^{(k)}\right)^*$领域内强凸函数且二次可导. 令$\hat{F}(\cdot;\beta^{(k)})$是残差项的经验累积分布函数, 则

$$\hat{Q}\left(\beta^{(k)}\right)=\int_{-\infty}^{\infty}\rho_\tau(\mu)d\,\hat{F}\left(\mu;\beta^{(k)}\right). \tag{4.1}$$

因为$\hat{F}(\cdot;\beta^{(k)})$是不连续的, 所以$\hat{Q}\left(\beta^{(k)}\right)$与$\rho_\tau(\cdot)$拥有相同的平滑度. 给定残差$r_i\left(\beta^{(k)}\right)=y_i^{(k)}-x_i^{(k)}\beta^{(k)}$与平滑参数$h=h_n>0$, 令$\hat{F}_h(\cdot;\beta^{(k)})$为 Parzen-Rosenblatt 核密度估计[20-21]的分布函数, 即

$$\hat{F}_h\left(\mu;\beta^{(k)}\right)=\int_{-\infty}^{\mu}\hat{f}_h(t;\beta^{(k)})dt,$$

其中$\hat{f}_h\left(t;\beta^{(k)}\right)=\frac{1}{n}\sum_{i=1}^{n}K_h(t-r_i\left(\beta^{(k)}\right))$, $K_h(\mu):=\frac{1}{h}K(\frac{\mu}{h})$, K为非负对称的核函数. 用$\hat{F}_h(\mu;\beta^{(k)})$去取代(4.1)式中的$\hat{F}\left(\mu;\beta^{(k)}\right)$, 则(4.1)式可变为

$$\hat{Q}_h\left(\beta^{(k)}\right):=\int_{-\infty}^{\infty}\rho_\tau(\mu)d\,\hat{F}_h\left(\mu;\beta^{(k)}\right)\frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{\mu}\rho_\tau(\mu)K\left(\frac{\mu+x_i^{(k)}\beta^{(k)}-y_i^{(k)}}{h}\right)d\mu. \tag{4.2}$$

在给定核函数$K(\mu)$与$h>0$的情况下(4.2)式还可改写为

$$\hat{Q}_h\left(\beta^{(k)}\right)=\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_h\left(y_i^{(k)}-x_i^{(k)}\beta^{(k)}\right), \tag{4.3}$$

其中

$$\mathcal{L}_h(\mu) = (\rho_\tau * K_h)(\mu) = \int_{-\infty}^{\infty} \rho_\tau(v) K_h(v-\mu) dv. \tag{4.4}$$

(4.4)式中的$*$代表卷积算子.

(4.2)式中的核函数$K(\mu)$有多种选择, 在后续的文章中考虑$K(\mu)$为高斯核函数的情况.

4.1.1 基于 LASSO 惩罚函数的分层卷积平滑分位数回归模型

为了使(4.2)式可以在高维情况下进行估计, 考虑对其施加 LASSO 惩罚函数, 模型中的系数通过以下目标函数求解:

$$\left(\hat{v}, \hat{\delta}^{(1)}, \cdots, \hat{\delta}^{(K)}\right) \in argmin\left\{\sum_{k=1}^{K} \hat{Q}_h\left(v+\delta^{(k)}\right) + \lambda_1 \|v\|_1 + \sum_{k=1}^{K} \lambda_2^{(k)} \left\|\delta^{(k)}\right\|_1\right\}. \tag{4.5}$$

为了方便对(4.5)式进行估计, 对(4.5)式做类似(2.3)-(2.4)式的变换, 令$\theta^{(k)} = \frac{\lambda_2^{(k)}}{\lambda_1}$

$$X^*_{n\times(K+1)p} = \begin{bmatrix} x^{(1)} & \frac{x^{(1)}}{\theta^{(1)}} & \cdots & 0_{n_k\times p} \\ \vdots & \vdots & \ddots & \vdots \\ x^{(K)} & 0_{n_k\times p} & \cdots & \frac{x^{(K)}}{\theta^{(K)}} \end{bmatrix},$$

$$\beta = (v^T, \left(\theta^{(1)}\delta^{(1)}\right)^T, \cdots, \left(\theta^{(K)}\delta^{(K)}\right)^T)^T.$$

则(4.5)式可变为

$$\hat{\beta} \in argmin\{\hat{Q}_h(\beta) + \lambda_1 \|\beta\|_1\}, \tag{4.6}$$

其中$\hat{Q}_h(\beta) \coloneqq \int_{-\infty}^{\infty} \rho_\tau(\mu) d\, \hat{F}_h\left(\mu; \beta^{(k)}\right) = \frac{1}{nh}\sum_{i=1}^{n} \int_{-\infty}^{\mu} \rho_\tau(\mu) K\left(\frac{\mu + X_i^*\beta - Y_i}{h}\right) d\mu.$

4.1.2 回归系数估计方法

为了对(11)式进行估计, 考虑如下迭代方法

$$\hat{\beta}^{(l)} \in argmin\{\hat{Q}_h(\beta) + q'_\lambda\left(\left|\hat{\beta}^{(l-1)}\right|\right)|\beta|\}, \tag{4.7}$$

其中$l = 1,2,\cdots$, 表示迭代次数, (4.7)式从$\hat{\beta}^{(0)}$开始迭代, $q_\lambda(\cdot)$为惩罚函数.

令$\hat{\beta}^{(0)} = \mathbf{0}, q'_\lambda\left(\left|\hat{\beta}_j^{(0)}\right|\right) = \lambda, \lambda_j^{(l-1)} = q'_\lambda\left(\left|\hat{\beta}_j^{(l-1)}\right|\right)$, 则(4.7)式可以变为

$$\hat{\beta}^{(l)} \in argmin\left\{\hat{Q}_h(\beta) + \left\|\boldsymbol{\lambda}^{(l-1)} \circ \beta\right\|_1\right\}, \tag{4.8}$$

其中$\boldsymbol{\lambda}$为由$\lambda_j \geq 0$构成的向量, $\circ$表示哈达玛积.

现在将$K(\mu)$为高斯核函数的情况带入(4.8)式, 使用 Gu 等(2018)提出的基于分位数回归的 ADMM 算法求解如下式子:

$$\begin{aligned} &minmize\left\{\hat{Q}_h(\beta) + \left\|\boldsymbol{\lambda}^{(l-1)}\beta\right\|_1\right\} \\ &s.t.\boldsymbol{r} = Y - X^*\beta, \end{aligned} \tag{4.9}$$

其中$\boldsymbol{r} = (r_1, \cdots, r_n)^T, r_i = Y_i - X_i^*\beta$, 则(4.9)式的增广拉格朗日形式如下:

$$\mathcal{L}_\rho(\beta, \boldsymbol{r}, \eta) = \hat{Q}_h(\beta) + \left\|\boldsymbol{\lambda}^{(l-1)}\beta\right\|_1 + \langle \eta, \boldsymbol{r} - Y_i + X_i^*\beta\rangle + \frac{\rho}{2}\|\boldsymbol{r} - Y_i + X_i^*\beta\|_2^2. \tag{4.10}$$

其中η是拉格朗日乘子, ρ是 ADMM 算法的参数.

ADMM 算法通过更新$\beta, \boldsymbol{r}, \eta$这三个参数进行迭代计算, 其详细迭代过程如下:

第一步, 令$\hat{\beta}^{(0)} = \hat{r}^{(0)} = \hat{\eta}^{(0)} = \mathbf{0}$, 并设置收敛率$\varepsilon$.

第二步, 通过下式更新β

$$\hat{\beta}^{(t)} = argmin\{\frac{\rho}{2}\left\|Y - \hat{r}^{(t-1)} - \frac{1}{\sqrt{\rho}}\hat{\eta}^{(t-1)} - X^*\beta\right\|_2^2 + \left\|\boldsymbol{\lambda}^{(l-1)}\beta\right\|_1\}.$$

第三步, 通过求解下式更新r_i

$$\tau - \Phi\left(\frac{-r_i}{h}\right) + \hat{\eta}_i^{(t-1)} + \rho\left(r_i - Y_i + \langle X_i^*, \hat{\beta}^{(t)}\rangle\right) = 0,$$

其中τ为分位数参数.

第四步, 通过下式更新η

$$\hat{\eta}^{(t)} = \hat{\eta}^{(t-1)} + \rho(\hat{r}^{(t)} - Y + X^*\hat{\beta}^{(t)}).$$

第五步, 若$\left\|\hat{\beta}^{(t)} - \hat{\beta}^{(t-1)}\right\|_2 \leq \varepsilon$, 则停止上述过程, 否则继续.

在本章中的惩罚函数项, 可以选取多种惩罚函数进行降维处理, 但是为了与分层线性回归模型进行比较, 选取$q_\lambda(\cdot)$为 LASSO 惩罚函数.

4.1.3 惩罚参数估计方法

现有分层分位数回归文献, 在估计(4.6)式的惩罚参数时, 使用网格法与 CV 方法这两种不同的算法对参数$\theta^{(k)}$与惩罚参数λ_1进行估计. 本文决定使用 R 语言

中的 sparsematrix 函数通过构造λ_1与$\lambda_2^{(k)}$的稀疏矩阵得到X^*. 在X^*的基础上, 基于(4.6)式, 仅使用 CV 方法就可以对惩罚参数λ_1、参数$\theta^{(k)}$进行求解. 在蒙特卡洛模拟与实例分析中也可以发现, 通过卷积平滑处理与上述惩罚参数估计方法得到的分层卷积平滑分位数回归模型在估计精度与估计效率方面都要优于分层分位数回归模型.

4.2 蒙特卡洛模拟

本节中, 使用蒙特卡洛模拟来研究所提方法的效果, 该模拟方法与第 3 章中的模拟方法类似, 只有在误差项构造、k的取值与部分$\beta_i^{(k)}$的构造出现些许不同. 数值模拟参数设置如下：$k=10, p=50, n_k \in \{40,50,100\}$, 观测矩阵$X \sim N(0,\Sigma)$, 其中$\Sigma$是$p \times p$维的 T 型矩阵, $\Sigma=(0.5^{|i-j|})_{1\le i,j\le p}$, 其中多个$n_k$的值是为了研究$p/n_k$的比值对估计精度的影响. 接下来开始构造$\beta^{(k)}$, 首先从$p$个维度中选择 20 个维度, 记作集合$P_0, P_0 \in \{1,2,\cdots,p\}$, 同时令$i \in \{1,2,\cdots,p\}$, 当$i \notin P_0$时, $\beta_i^{(k)}=0$, 当$i \in P_0$时, 再从P_0中抽取 10 个元素, 组成集合P_1, 当$i \in P_0$且$i \notin P_1$时, $\beta^{(k)}$分两种情况构造, 当$k<3$时, $\beta_i^{(k)}=1$, 当$\mathrm{k}>3$时, $\beta_i^{(k)}=1+\gamma^k$；当$i \in P_1$时, $\beta^{(k)}$也分两种情况构造, 当$k<3$时, $\beta_i^{(k)}=1+\gamma^k$, 当$k>3$时, $\beta_i^{(k)}=1$, 上述过程中令$\gamma^k=0.1\times K^{1/2}$. 通过上述过程, 构造共性回归系数是 1、个性回归系数是 0 或γ^k. 记$\beta^\circ=(\beta^{(1)},\cdots,\beta^{(10)})$, $\hat{\beta}^\circ$为其估计值. 随机误差项$\varepsilon^{(k)}$由如下 4 种方法构造.

(1) $\epsilon^{(k)} \backsim t(1.5)$.

(2) $\epsilon^{(k)} \backsim N(0,1)$.

(3) $\epsilon^{(k)} \backsim C(1,0)$.

(4) $\epsilon^{(k)} \backsim N(0,1)$, 随机选出 40%, 其中的一半取 8, 另一半取-8, 模拟异常值情况.

因为第 4 章的重心是两种分位数模型的比较, 所以在误差项的构造时, 选取

$C(1,0)$(柯西分布)、$t(1.5)$(t分布)这样具有尖峰与厚尾特征的分布.

选取 TPR、FPR、ℓ_2这 3 个评价指标来衡量模型的估计性能, 计算方法如下：

$$\mathrm{TPR}=\frac{\#(i:\widehat{\beta}_i\neq 0 \text{ 且 } \beta_i^*\neq 0)}{\#(i:\beta_i^*\neq 0)},$$

$$\mathrm{FPR}=\frac{\#(i:\widehat{\beta}_i\neq 0 \text{ 且 } \beta_i^*= 0)}{\#(i:\beta_i^*= 0)},$$

$$\ell_2=\left\|\hat{\beta}_i-\beta_i^*\right\|_2.$$

TPR、FPR 分别表示正确估计非 0 回归系数的比例与错误估计非 0 回归系数的比例. TPR 越大、FPR 越小表示模型的估计精度越高. ℓ_2用来衡量模型回归系数的误差, ℓ_2越小估计精度越高. p/n_k的不同比值是为了探究维度p对模型估计的影响. 选取分层最小二乘回归模型、分层分位数回归模型作为对比模型. 为了方便表示, 分层卷积平滑分位数回归模型记为SQR, 分层最小二乘回归模型记为LR, 分层分位数回归模型记为QR. 为了与LR模型进行对比, QR模型与SQR模型均选择 0.5 分位数进行回归模拟. 惩罚参数通过 10 折 CV 方法进行估计.

表 4.1 3 种模型的模拟结果

p/n_k	误差项	模型	TPR	FPR	ℓ_2
>1	(1)	LR	0.976	0.208	7.033
		QR	1.000	0.103	2.575
		SQR	1.000	0.068	2.077
	(2)	LR	1.000	0.447	1.576
		QR	1.000	0.097	1.954
		SQR	1.000	0.126	1.641
	(3)	LR	0.486	0.084	12.400
		QR	1.000	0.073	2.957
		SQR	1.000	0.072	2.550
	(4)	LR	1.000	0.431	4.010
		QR	1.000	0.128	3.072
		SQR	1.000	0.101	2.357
=1	(1)	LR	0.998	0.164	5.756
		QR	1.000	0.082	2.291
		SQR	1.000	0.063	1.746
	(2)	LR	1.000	0.436	1.172
		QR	1.000	0.084	1.752
		SQR	1.000	0.136	1.389
	(3)	LR	0.716	0.086	11.738

续表 4.1

p/n_k	误差项	模型	TPR	FPR	ℓ_2
		QR	1.000	0.079	2.253
		SQR	1.000	0.071	1.992
	(4)	LR	1.000	0.354	3.674
		QR	1.000	0.107	2.094
		SQR	1.000	0.094	2.041
<1	(1)	LR	1.000	0.107	4.822
		QR	1.000	0.062	2.175
		SQR	1.000	0.062	1.719
	(2)	LR	1.000	0.403	1.254
		QR	1.000	0.058	1.411
		SQR	1.000	0.084	1.327
	(3)	LR	0.733	0.047	11.004
		QR	1.000	0.091	1.794
		SQR	1.000	0.060	1.477
	(4)	LR	1.000	0.247	2.936
		QR	1.000	0.104	1.522
		SQR	1.000	0.098	1.525

观察表 4.1 可以发现, 随着p/n_k的减小, 针对同一误差项, 3 种模型的估计精度都会提升：TPR 呈现出增加的趋势, FPR 与ℓ_2呈现出减小的趋势.

LR 模型在误差项服从标准正态分布时表最好, ℓ_2要小于 SQR 模型与 QR 模型, 但是 FPR 要远大于 SQR 模型与 QR 模型, 表明相较于 SQR 模型与 QR 模型, LR 模型倾向于估计较多的非 0 系数. 同时 LR 模型在误差项服从标准柯西分布与$t(1.5)$分布时表现最差, ℓ_2远大于 SQR 模型与 QR 模型且 TPR 也要小于 1, 表明 LR 模型相较于 QR 模型与 SQR 模型, 难以正确的估计模型中的非 0 系数, 且估计精度较差. LR 在 4 种误差项的情况下均拥有最大的 FPR, 表明 LR 模型相较于 QR 模型与 SQR 模型, 容易错误的估计非 0 系数.

QR 模型在$p/n_k > 1$且误差项存在异常值的情况下, ℓ_2要小于 SQR 模型但两者相差不大. 在其余 3 种误差项与误差项存在异常值且满足$p/n_k < 1$、$p/n_k = 1$情况下, QR 模型的ℓ_2要大于 SQR 模型. 同时可以发现, QR 模型与 SQR 模型拥有相同的 TPR, 表明这 2 种模型均可以正确的估计出非 0 回归系数. 针对数值模拟的 4 种误差项分布, 在 FPR 这一指标下 SQR 模型要优于 QR 模型, 表明相较于

SQR 模型, QR 模型倾向于估计较多的非 0 系数.

SQR 模型除了在误差项存在异常值与误差项服从标准正态分布的情况下, 在 3 种模型中均拥有最小的 FPR 与ℓ_2, 表明 SQR 模型无论是在估计精度还是非 0 系数的估计方面, 均要优于 LR 模型与 QR 模型.

综上, 当数据误差项不满足正态分布时, QR 模型与 SQR 模型要优于 LR 模型, 针对其余 3 种误差项, SQR 模型在大多数情况下, 在 3 种衡量指标下表现最优. p/n_k的比值会对 3 种模型的估计精度都产生影响, 随着p/n_k的增加, 3 种模型的估计精度都会下降.

下面从模型估计效率方面对比与探究分层线性回归模型、分层分位数回归模型与分层卷积分位数回归模型.

表 4.2 3 种模型的估计效率(单位：秒)

p/n_k	误差项	模型	速度
>1	(1)	LR	0.25
		QR	388.65
		SQR	10.45
	(2)	LR	0.14
		QR	353.07
		SQR	6.80
	(3)	LR	0.40
		QR	334.41
		SQR	16.52
	(4)	LR	0.22
		QR	345.64
		SQR	10.78
=1	(1)	LR	0.36
		QR	510.39
		SQR	11.22
	(2)	LR	0.18
		QR	420.04
		SQR	7.49
	(3)	LR	0.64

续表 4.2

p/n_k	误差项	模型	速度
		QR	364.89
		SQR	18.28
	(4)	LR	0.28
		QR	398.78
		SQR	12.19
<1	(1)	LR	0.55
		QR	1525.48
		SQR	17.39
	(2)	LR	0.24
		QR	1452.41
		SQR	7.78
	(3)	LR	0.77
		QR	1448.40
		SQR	24.34
	(4)	LR	0.36
		QR	1404.22
		SQR	16.19

通过观察表 4.2 可以发现, LR 模型的估计速度最快, SQR 模型次之, QR 模型最慢. 在任意p/n_k的情况下, QR 模型在误差项服从$t(1.5)$分布时估计效率最差, LR 模型与 SQR 模型在误差项服从标准柯西分布的情况下估计效率最低. 随着n_k的增加, 3 种模型的估计效率呈现出降低的趋势. 相较于 SQR 模型与 LR 模型, QR 模型对n_k的变化最为敏感, 随着n_k的增加, QR 模型估计用时增加幅度最大, 几乎呈现出翻倍的趋势.

综上所述, SQR 模型在估计精度方面要明显优于 QR 模型与 LR 模型, 在估计效率方面, SQR 模型虽然略逊于 LR 模型, 但要明显优于 QR 模型.

4.3 实例分析

在这一节中, 同样选取 Kouno 等(2013)论文中的数据集来验证 SQR 模型在实际数据集中的表现. 该数据集的尖峰厚尾的特性在第 3 章中进行了论述, 所

以在这里不在赘述，直接看 SQR 模型、QR 模型与 LR 模型估计的结果.

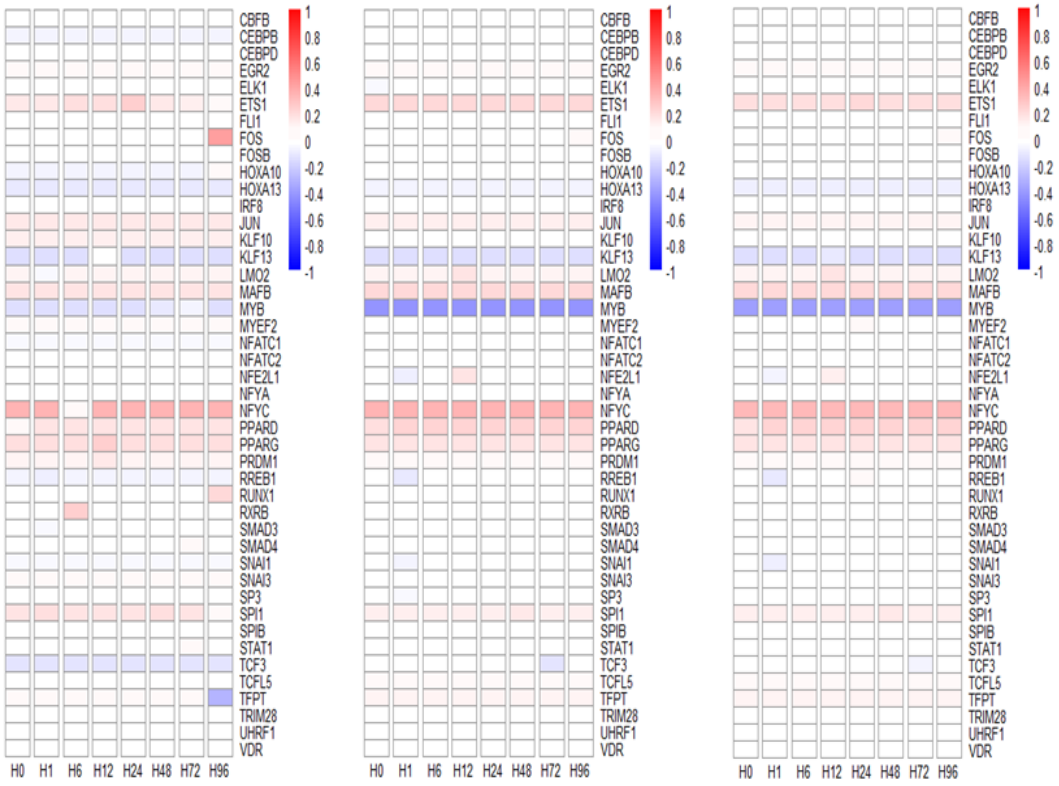

图 4.1　3 种模型系数的热力图(从左到右分别是 LR 模型、QR 模型、SQR 模型)

通过图 4.1 可以发现，当选择与 BCL6 转录因子表达呈现出正相关的转录因子时, NFYC、PPARD、PPARG、MAFB、JUN、ETS1、SPI1 这 7 个转录因子在 3 种模型中均与 BCL6 转录因子呈现出较为明显的正向相关关系, EGR2、LMO2 等转录因子与 BCL6 转录因子呈现出不太明显的正相关关系. Kouno 等(2013)研究发现, PPARG、MAFB、ETS1、SPI1、JUN、EGR2、KLF10 这 7 个转录因子对 BCL6 转录因子的表达起正向作用. 对比图 2 可以发现 LR 模型多选择了 PRDM1、MYEF2、TFPT、NFYC 这 4 个转录因子，同时 LR 模型在 H96 时 BCL6 与 ETS1 的正相关性过小，这与 Kouno 等(2013)研究指出的 BCL6 与 ETS1 的强相关性的研究结果相违背. SQR 模型与 QR 模型仅多选择了 TCFL5、NFYC 这 2 个转录因子，但是却漏选了 KLF10 这个转录因子.

当选择与 BCL6 转录因子表达呈现出负相关的转录因子时, SQR 模型与 QR 模型仅选择了 MYB、KLF13、HOXA13 这 3 个转录因子且 MYB 转录因子的系数最大. 而 LR 模型在 MYB、KLF13、HOXA13 这 3 个转录因子的基础上还多选

择了 CEBPB、HOXA10、RREB1、TCF3 这 4 个转录因子, 同时 MYB 转录因子的系数与 KLF13、HOXA13 等转录因子的系数相比也没有明显的区别. Kouno 等(2013)研究指出, BCL6 转录因子的表达仅与 MYB 转录因子呈现出明显的负相关关系. LR 模型选择了过多没有关系的转录因子, SQR 模型与 QR 模型表现更好.

SQR 模型与 QR 模型虽然选择出的与 BCL6 转录因子表达有关的转录因子大致相同, 但 SQR 模型相较于 QR 模型, 诸如 SP3(H1)、ELK1(H0)、TCF3(H72)这样单独出现的干扰系数较少, 模型更为稳健.

表 4.3　3 种模型在实际数据中的估计效率(单位：秒)

模型	LR	QR	SQR
速度	4.32	282.37	15.06

表 4.3 可以很明显的看出 SQR 模型与 LR 模型的估计效率明显优于 QR 模型, SOR 模型的估计速度略慢于 QR 模型.

综上所述, 在实际应用中 SQR 模型更为精确与稳健, 也拥有不错的估计效率, 相较于 QR 模型与 LR 模型具有更好的实用价值.

5 研究结论与展望

5.1 研究结论

对分层数据进行建模有利于发掘出数据背后的规律, 得到估计更加精确的模型. 但是, 目前该方向的研究考虑数据是完全无缺失的假设过于理想化, 不太符合现实生活中因为各种问题数据出现缺失的情况. 同时, 因为分位数回归模型损失函数不可微这一缺点, 会导致分层分位数回归模型估计精度降低, 估计效率变差. 因此, 本文针对上述的两个缺点, 对分层分位数回归模型进行拓展与改进, 得出了如下结论：

(1) 基于逆概率加权方法与变量选择方法研究了响应变量随机缺失与异方差情况下分层分位数回归模型参数估计与变量选择问题, 并在一定条件下证明了所提估计的渐近正态性与 Oracle 性质. 通过数值模型与实例分析可以发现, 所提估计方法相较于对随机缺失的响应变量不加处理的方法, 其估计更加精确. 同时所提估计方法在误差项不满足正态分布时要优于分层最小二乘回归模型.

(2) 基于卷积平滑方法来改进分层分位数回归模型, 提出了分层卷积平滑分位数回归模型. 该模型结合卷积平滑方法、分位数回归模型与正则化惩罚技术, 在针对响应变量分层的数据建模时, 不仅在估计过程中降低了变量维度、提升了估计精度, 同时也减轻了计算负担, 提升了估计效率. 数值模拟发现分层卷积平滑分位数回归模型在处理具有尖峰厚尾特征的数据时要优于分层分位数回归模型与分层最小二乘回归模型; 分层卷积平滑分位数回归模型的估计效率也要远高于分层分位数回归模型. 实例分析的结果也体现了分层卷积平滑分位数回归模型的优越性.

5.2 研究展望

(1) 分层分位数回归模型在模拟时必须提前设置其层数, 如果不能提前获知数据的层数, 将无法进行后续的分析研究工作. 后续的研究希望可以改善这一缺陷.

(2) 本文只考虑了响应变量随机缺失的情况, 但在实际的生活中, 协变量也

会出现缺失的情况. 在生物医学领域, 删失与截断的数据同样也很常见, 在后续的研究中, 希望可以考虑上述的情况, 对模型进行推广研究.

(3) 目前关于分层回归模型的研究多集中于生物医学领域, 但在社会学、经济学等领域中也会出现具有分层特性的数据且具有尖峰厚尾的特征. 在后续的研究中, 希望将本问提出的模型拓展到上述领域.

参考文献

[1] Bai Y, Tian M, Tang M L, et al. Variable selection for ultra-high dimensional quantile regression with missing data and measurement error[J]. Statistical Methods in Medical Research, 2020, 30(1):129-150.

[2] Ballout N, Garcia C, Viallon V. Sparse estimation for case-control studies with multiple subtypes of cases[J]. arXiv preprint arXiv:1901.01583, 2019.

[3] Brockwell P J, Davis R A, Berger J O, et al. Time series: theory and methods[M]. Springer-Verlag, 2015.

[4] Christou E, Akritas M G. Variable selection in heteroscedastic single-index quantile regression[J]. Communications in Statistics-Theory and Methods, 2018, 47(24):6019-6033.

[5] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties[J]. Journal of the American statistical Association, 2001, 96(456): 1348-1360.

[6] Fan J, Lv J. Sure Independence Screening for Ultra-High Dimensional Feature Space[J]. Journal of the Royal Statistical Society, 2008, 70(5):849-911

[7] Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2008, 70(5): 849-911.

[8] Fan J, Xue L, Zou H. Strong oracle optimality of folded concave penalized estimation[J]. Annals of statistics, 2014, 42(3): 819-849.

[9] Fan J, Xue L, Zou H. Strong oracle optimality of folded concave penalized estimation[J]. Annals of statistics, 2014, 42(3): 819-849.

[10] Fan R, Lee J H. Predictive quantile regressions under persistence and conditional heteroskedasticity[J]. Journal of Econometrics, 2019, 213(1):261-280.

[11] Galvao A F, Kato K. Smoothed quantile regression for panel data[J]. Journal of econometrics, 2016, 193(1): 92-112.

[12] Gertheiss J, Tutz G. Sparse modeling of categorial explanatory variables[J]. The Annals of Applied Statistics, 2010, 4(4): 2150-2180.

[13] Gu Y, Fan J, Kong L, et al. ADMM for high-dimensional sparse penalized quantile regression[J]. Technometrics, 2018, 60(3): 319-331.

[14] Han P, Kong L, Zhao J, et al. A general framework for quantile estimation with incomplete data[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2019, 81(2):305-333.

[15] Hautsch N, Kyj L M, Hautsch N. A blocking and regularization approach to high dimensional realized covariance estimation[J]. CFS Working Paper Series, 2009, 27(4):625-645.

[16] He X, Pan X, Tan K M, et al. Smoothed quantile regression with large-scale inference[J]. Journal of Econometrics, 2021.

[17] Hjort N L, Pollard D. Asymptotics for minimisers of convex processes[J]. arXiv preprint arXiv:1107.3806, 2011.

[18] Horowitz J L. Bootstrap methods for median regression models[J]. Econometrica, 1998, 66(6): 1327-1351.

[19] Horvitz D G, Thompson D J. A generalization of sampling without replacement from a finite universe[J]. Journal of the American statistical Association, 1952, 47(260):663-685.

[20] Hu J, Huang J, Qiu F. A group adaptive elastic-net approach for variable selection in high-dimensional linear regression[J]. Science China Mathematics, 2018, 61(1): 173-188.

[21] James Douglas Hamilton. Time Series Analysis[M]. Princeton University Press,1994.

[22] Knight K. Asymptotics for L-1 regression estimators under general conditions[J]. Annals of Statistics, 1997, 26(2):755-770.

[23] Koenker R, Bassett Jr G. Regression quantiles[J]. Econometrica: journal of the Econometric Society, 1978, 46(1):33-50.

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

[25] Koenker, R. Quantile Regression[M]. New York: Cambridge University Press, 2005.

[26] Kouno T, Hoon M D, Mar J C, et al. Temporal dynamics and transcriptional control using single-cell gene expression analysis[J]. Genome Biology, 2013, 14(10):R118.

[27] Little R, Rubin D B. Statistical analysis with missing data[J]. Technometrics, 2002, 45(4): 364-365

[28] Ollier E, Viallon V. Regression modelling on stratified data with the lasso[J]. Biometrika, 2017, 104(1):83-96.

[29] Parzen E. On estimation of a probability density function and mode[J]. The annals of mathematical statistics, 1962, 33(3): 1065-1076.

[30] Peng B, Wang L. An iterative coordinate descent algorithm for high-dimensional nonconvex penalized quantile regression[J]. Journal of Computational and Graphical Statistics, 2015, 24(3): 676-694.

[31] Rosenblatt M. A central limit theorem and a strong mixing condition[J]. Proceedings of the national Academy of Sciences, 1956, 42(1): 43-47.

[32] Rothman A J, Levina E, Zhu J. Generalized thresholding of large covariance matrices[J]. Journal of the American Statistical Association, 2009, 104(485): 177-186.

[33] Rubin D B. Multiple imputation for nonresponse in surveys[M]. John Wiley & Sons, 2004.

[34] Sherwood B, Wang L, Zhou X H. Weighted quantile regression for analyzing health care cost data with missing covariates[J]. Statistics in medicine, 2013, 32(28):4967-4979.

[35] Sherwood B. Variable selection for additive partial linear quantile regression with missing covariates[J]. Journal of Multivariate Analysis, 2016, 152:206-223.

[36] Tan K M, Wang L, Zhou W X. High-dimensional quantile regression: Convolution smoothing and concave regularization[J]. Journal of the Royal Statistical Society: Series B (Methodological), 2022, 84(1): 205-233.

[37] Tibshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 267-288.

[38] Torrenté de L, Zimmerman S, Suzuki M, et al. The shape of gene expression distributions matter: how incorporating distribution shape improves the

interpretation of cancer transcriptomic data[J]. BMC bioinformatics, 2020, 21(21): 1-18.

[39] Wang H, Leng C. A note on adaptive group lasso[J]. Computational statistics & data analysis, 2008, 52(12): 5277-5286.

[40] Wang L, Wu Y, Li R. Quantile regression for analyzing heterogeneity in ultra-high dimension[J]. Journal of the American Statistical Association, 2012, 107(497): 214-222.

[41] Wu Y, Ma Y, Yin G. Smoothed and corrected score approach to censored quantile regression with measurement errors[J]. Journal of the American Statistical Association, 2015, 110(512): 1670-1683.

[42] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006, 68(1): 49-67.

[43] Zhao J, Shao J. Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data[J]. Journal of the American Statistical Association, 2015, 110(512): 1577-1590.

[44] Zhao P, Tang X. Imputation based statistical inference for partially linear quantile regression models with missing responses[J]. Metrika, 2016, 79(8):991-1009.

[45] Zhao, Xin P. Quantile regression for partially linear models with missing responses at random[J]. Applied Mechanics and Materials, 2015, 727-728: 1013-1016.

[46] Zheng Y, Zhu Q, Li G, et al. Hybrid quantile regression estimation for time series models with conditional heteroscedasticity[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2018, 80(5):975-993.

[47] Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models[J]. Annals of statistics, 2008, 36(4): 1509-1533.

[48] Zou H. The adaptive lasso and its oracle properties[J]. Journal of the American statistical association, 2006, 101(476): 1418-1429.

[49] 刘栋, 杨冬梅, 何勇, 张新生. 分组数据分位数回归模型的变量选择和估计[J].应用数学学报, 2021, 44(05):722-739.

[50]刘庆丰, 郑苗苗. 带有完全随机缺失协变量的广义线性模型平均方法研究[J]. 数量经济研究, 2020, 11(04):25-40.

[51]于力超. 非随机缺失机制下基于模型的参数似然估计方法研究[J]. 数理统计与管理, 2019, 38(6):977-985.

致 谢

时光飞逝, 转眼间三年的研究生生活即将接近尾声. 这三年时光是我人生中的重要阶段. 在硕士学习和生活中我遇到了许多对我帮助巨大的人. 首先, 我要感谢我的导师, 在生活和工作中, 他的建议和鼓励让我获得了很大的成长；在专业领域, 他向我介绍了统计学的各种前沿与理论知识, 并且对我进行了有关统计思维方式的教育, 我非常感谢他在过去两年中为我付出的所有时间和精力, 并且为成为的学生而感到自豪. 其次, 我要感谢同门的师兄师姐师弟师妹们, 他们在生活和学习上都给予了我很大的照顾与帮助；我还要感谢统计学院的所有老师, 老师们在课堂上的淳淳教诲、耐心的指导和毫无保留的传道解惑, 让我学到了许多专业的统计学知识和专业技能. 最后, 我要特别感谢我的父母, 感谢他们一直以来的支持与爱护, 是他们给予我力量, 让我面对困难时乐观而坚强, 激励着我不断前进, 完成我的求学之路.

回首研究生三年的学习时光, 因为疫情原因, 经历了封校、宿舍封控等一系列事情, 但是还是对段家滩 496 号有很多不舍, 马上就到了与母校分别的日子. 希望接下来我能把在学校学习的知识和技能运用于工 作中, 不辜负老师的教导和父母的期望.

附 录

给出在0.25以及0.75分位数的情况下,QR的估计结果以及异方差情况下QR的估计结果. 另外的, 为了更加直观的看出分层的方法的优势, 我们给出了每层分别估计的拟合结果. 为了方便我们将每层分别估计的方法记作 DQR. 之后还给出了异方差情况下的拟合结果以及 DQR 拟合结果. 最后还给出了异方差情况下各模型系数估计散点图.

表 1 $p/n_k = 1$时模型估计结果

τ	模型	误差项	缺失率	TPR	FPR	ℓ_2
0.25	QR	(1)	8%	1.000	0.043	2.087
	QR		30%	1.000	0.183	2.669
	QR		67%	1.000	0.297	5.260
	QR	(2)	8%	1.000	0.020	2.234
	QR		30%	1.000	0.047	2.400
	QR		67%	1.000	0.250	4.404
	QR	(3)	8%	1.000	0.110	1.921
	QR		30%	1.000	0.170	3.051
	QR		67%	1.000	0.463	8.110
0.75	QR	(1)	8%	1.000	0.123	1.636
	QR		30%	1.000	0.253	2.179
	QR		67%	1.000	0.337	3.756
	QR	(2)	8%	1.000	0.083	1.828
	QR		30%	1.000	0.103	1.956
	QR		67%	1.000	0.226	3.425
	QR	(3)	8%	1.000	0.103	1.812
	QR		30%	1.000	0.087	2.658
	QR		67%	0.968	0.107	6.395

观察表 1 的结果可以看出, 在$\tau = 0.25$时, 基于 QR 方法的 TPR 均为 1, 说明该方法能正确估计非 0 回归系数. 而在同一误差项下, 随着缺失率的提高, 基于 QR 方法的 FPR 以及估计误差ℓ_2均逐渐变大, 因此可以认为缺失率越高, 该方法对于模型估计的精度越低.

在$\tau = 0.75$时, 只有当误差项为(3)并且缺失率为 67%时, 基于 QR 方法的 TPR 为 0.968, 接近于 1, 其余情况下均为 1, 因此可以认为该方法能正确估计非 0 回归系数. 同理, 当误差项为(1), (2)时, 随着缺失率的提高, 基于 QR 方法的

FPR 以及估计误差ℓ_2均逐渐变大, 因此可以认为缺失率越高, 该方法对于模型估计的精度越低. 但是当误差项为(3)时, 基于 QR 方法的 FPR 相差不大, 但是基于 QR 方法的估计误差ℓ_2逐渐变大, 因此可以得出与上述情况一样的结论. 即随着缺失率的增加, 该方法对于模型估计的精度会降低.

综上, 在$p/n_k = 1$时, $\tau = 0.25$以及$\tau = 0.75$的情况下, 对于三种不同的误差项, 随着缺失率的提高, 基于 QR 方法对于模型估计的精度均会下降.

表 2 $p/n_k < 1$时模型估计结果

τ	模型	误差项	缺失率	TPR	FPR	ℓ_2
0.25	QR	(1)	8%	1.000	0.057	1.485
	QR		30%	1.000	0.140	1.628
	QR		67%	1.000	0.250	3.916
	QR	(2)	8%	1.000	0.047	1.323
	QR		30%	1.000	0.063	1.474
	QR		67%	1.000	0.120	3.261
	QR	(3)	8%	1.000	0.103	1.525
	QR		30%	1.000	0.133	2.323
	QR		67%	1.000	0.340	7.471
0.75	QR	(1)	8%	1.000	0.103	1.319
	QR		30%	1.000	0.113	1.744
	QR		67%	1.000	0.190	3.609
	QR	(2)	8%	1.000	0.053	1.330
	QR		30%	1.000	0.113	1.496
	QR		67%	1.000	0.180	2.674
	QR	(3)	8%	1.000	0.040	1.661
	QR		30%	1.000	0.110	1.669
	QR		67%	1.000	0.230	3.383

表 3 $p/n_k > 1$时模型估计结果

τ	模型	误差项	缺失率	TPR	FPR	ℓ_2
0.25	QR	(1)	8%	1.000	0.087	2.349
	QR		30%	1.000	0.177	3.491
	QR		67%	1.000	0.247	6.880
	QR	(2)	8%	1.000	0.177	1.890
	QR		30%	1.000	0.327	2.778
	QR		67%	1.000	0.330	4.528
	QR	(3)	8%	1.000	0.150	3.553
	QR		30%	1.000	0.253	5.957
	QR		67%	0.965	0.297	9.887

续表 3

τ	模型	误差项	缺失率	TPR	FPR	ℓ_2
0.75	QR	(1)	8%	1.000	0.087	2.405
	QR		30%	1.000	0.107	3.723
	QR		67%	1.000	0.130	5.085
	QR	(2)	8%	1.000	0.067	2.004
	QR		30%	1.000	0.123	2.918
	QR		67%	1.000	0.207	4.342
	QR	(3)	8%	1.000	0.043	2.268
	QR		30%	1.000	0.130	3.764
	QR		67%	0.908	0.290	8.975

观察表 2 以及表 3, 可以得出与表 1 相同的结论.

表 4 异方差情况下模型估计结果

p/n_k	τ	模型	缺失率	TPR	FPR	ℓ_2
$=1$	0.25	QR	8%	0.973	0.033	8.379
		QR	30%	0.875	0.051	10.525
		QR	67%	0.565	0.057	14.202
	0.75	QR	8%	0.900	0.053	9.513
		QR	30%	0.850	0.071	11.436
		QR	67%	0.745	0.093	15.922
>1	0.25	QR	8%	0.921	0.103	8.753
		QR	30%	0.845	0.183	8.971
		QR	67%	0.620	0.081	13.861
	0.75	QR	8%	0.975	0.067	5.851
		QR	30%	0.916	0.130	6.657
		QR	67%	0.795	0.120	11.321
<1	0.25	QR	8%	0.985	0.057	5.941
		QR	30%	0.935	0.073	8.429
		QR	67%	0.705	0.123	10.836
	0.75	QR	8%	0.993	0.067	5.851
		QR	30%	0.958	0.130	6.657
		QR	67%	0.710	0.120	11.321

根据表 4 的结果可得, 在各个n_k, p以及τ的组合下, 随着缺失率的增加, 基于 QR 方法的 TPR 会降低, 基于 QR 方法的 FPR 以及估计误差ℓ_2会逐渐增加. 因此可以认为随着缺失率的提高, 基于 QR 方法对于模型估计的精度会下降. 并且对比表 1, 2, 3 可以看出, 异方差会降低模型估计的准确性.

表 5 每层分别估计的拟合结果

p/n_k	误差项	缺失率	TPR	FPR	ℓ_2
$=1$	(1)	8%	0.785	0.243	10.892
		30%	0.775	0.253	12.569
		67%	0.320	0.133	16.707
	(2)	8%	0.860	0.247	10.147
		30%	0.805	0.273	11.341
		67%	0.315	0.093	16.138
	(3)	8%	0.825	0.267	11.342
		30%	0.680	0.333	14.100
		67%	0.250	0.097	16.635
>1	(1)	8%	0.760	0.247	11.993
		30%	0.610	0.207	13.975
		67%	0.235	0.070	17.336
	(2)	8%	0.825	0.240	11.074
		30%	0.625	0.247	13.366
		67%	0.205	0.053	16.734
	(3)	8%	0.705	0.223	12.960
		30%	0.590	0.267	14.884
		67%	0.180	0.067	16.359
<1	(1)	8%	1.000	0.273	4.599
		30%	0.975	0.350	5.363
		67%	0.795	0.377	12.574
	(2)	8%	1.000	0.377	3.799
		30%	1.000	0.457	5.668
		67%	0.880	0.427	10.673
	(3)	8%	0.985	0.360	6.882
		30%	0.935	0.313	8.677
		67%	0.675	0.309	14.117

根据表 5 的结果, 可以得出结论, 当$p/n_k = 1$和$p/n_k > 1$时, 随着缺失率的增加, 每层分别估计出的 TPR 明显降低, 同时 FPR 也明显降低, 但是基于 LR 方法的估计误差ℓ_2增加, 根据图 1 的结果可以得到, 模型估计出的非 0 系数个数急剧的减小, 导致了 FPR 出现下降的情况.

而$p/n_k < 1$时, 可以认为随着缺失率的提高, 基于 QR 方法对于模型估计的精度会下降.

表 6 异方差情况下拟合结果

p/n_k	τ	模型	缺失率	TPR	FPR	ℓ_2
$=1$	0.25	QR	8%	0.915	0.033	8.379

续表 6

p/n_k	τ	模型	缺失率	TPR	FPR	ℓ_2
		QR	30%	0.710	0.020	12.525
		QR	67%	0.325	0.057	15.202
	0.75	QR	8%	0.900	0.053	9.513
		QR	30%	0.850	0.071	11.436
		QR	67%	0.745	0.093	15.922
>1	0.25	QR	8%	0.921	0.103	7.853
		QR	30%	0.845	0.183	8.971
		QR	67%	0.620	0.080	13.861
	0.75	QR	8%	0.975	0.067	5.851
		QR	30%	0.916	0.130	6.657
		QR	67%	0.795	0.120	11.321
<1	0.25	QR	8%	0.985	0.057	8.558
		QR	30%	0.935	0.073	10.429
		QR	67%	0.705	0.123	18.836
	0.75	QR	8%	0.993	0.067	5.851
		QR	30%	0.958	0.130	6.657
		QR	67%	0.710	0.120	11.321

根据表 6 的结果可得, 异方差的情况下, 在各个n_k, p以及τ的组合下, 随着缺失率的增加, 基于 QR 方法的 TPR 会降低, 基于 QR 方法的 FPR 以及估计误差ℓ_2会逐渐增加. 因此可以认为随着缺失率的提高, 基于 QR 方法对于模型拟合的效果会降低.

表 7 异方差情况下 DQR 拟合结果

p/n_k	τ	缺失率	TPR	FPR	ℓ_2
$=1$	0.25	8%	0.4100	0.1333	16.9785
		30%	0.3010	0.1267	16.2138
		67%	0.2100	0.1167	23.8692
	0.75	8%	0.3400	0.2067	15.4674
		30%	0.2450	0.1667	16.1279
		67%	0.1950	0.1333	20.8105
>1	0.25	8%	0.2650	0.1333	15.3001
		30%	0.1500	0.1000	16.1660
		67%	0.1250	0.0833	22.9300
	0.75	8%	0.2000	0.0933	16.1550
		30%	0.1300	0.0467	15.5296
		67%	0.1300	0.0667	18.7619

续表 7

p/n_k	τ	缺失率	TPR	FPR	ℓ_2
<1	0.25	8%	0.6350	0.1700	13.7264
		30%	0.5350	0.1267	16.9224
		67%	0.3150	0.1133	25.8331
	0.75	8%	0.6200	0.1767	12.5757
		30%	0.3700	0.1333	14.0278
		67%	0.2950	0.1033	22.9930

根据表 7 的结果可以看出，拟合结果的规律与表 5 的规律相同，但是在异方差的情况下，每层分别估计的 TPR 较低，因此拟合效果较差.

异方差情况下 QR、LR 与 DQR 模型的估计精度较低，可以通过如下的系数估计散点图更为直观的看出 3 种模型的优劣.

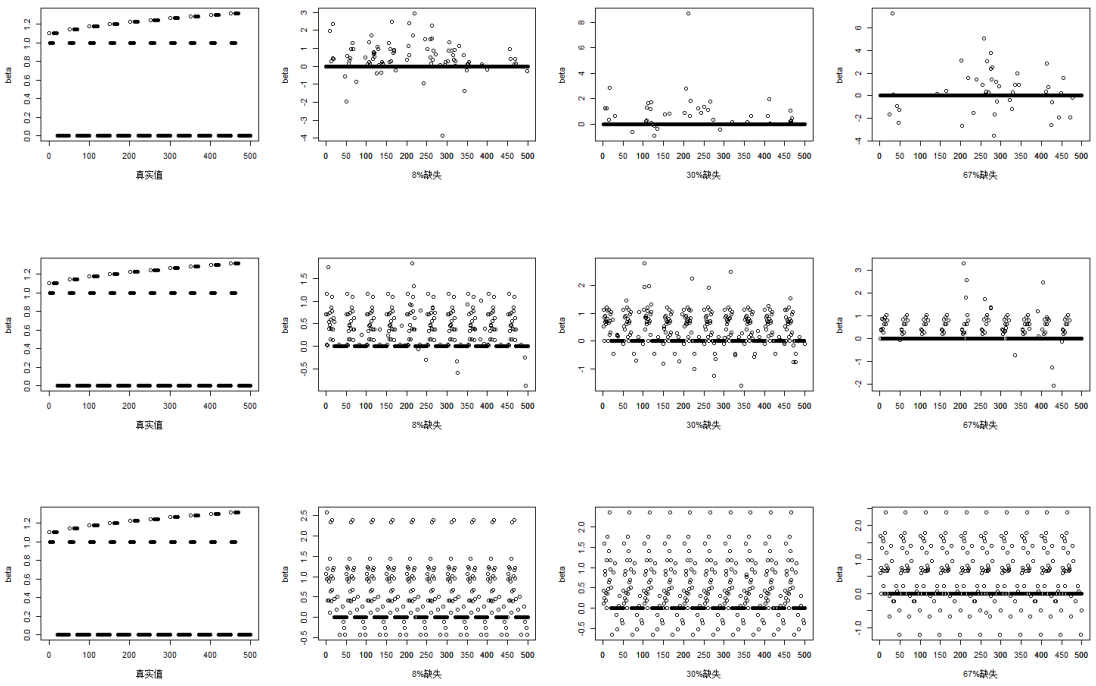

图 1 异方差情况下各模型系数估计散点图

图 1 从上往下给出了 DQR, QR, LR 三种方法在异方差的情况下对于模型系数估计的散点图. 从图中我们可以看出随着缺失率的提高，基于三种方法的非 0 系数个数均呈现下降的趋势. 并且根据系数真实值非零系数的情况，基于 QR 方

法的非 0 系数个数最接近于真实值, 而基于 DQR 方法的非零系数个数最少, 因此认为 QR 方法优于 DQR 方法和 LR 方法. 接下来看系数的真实值分布情况, 除了等于 0 的系数之外, 其余的非 0 系数大致分布于 1.0 到 1.3 之间. 显然根据估计值的范围来看, 基于 QR 方法的估计值相比于 DQR 与 LR 方法更接近于真实的情况. 综上, 在异方差的情况下, QR 方法要优于 DQR 与 LR 方法.