多节点分位数回归模型在睡眠时间和抑郁程度研究中的应用*

复旦大学公共卫生学院生物统计学教研室(200032)　潘璐璐　秦国友△

【提　要】 **目的**　介绍多节点分位数回归模型及其在睡眠时间和抑郁程度关联分析中的应用。**方法**　基于 NHANES 数据库,使用 R 语言软件 MultiKink 包拟合多节点分位数回归模型,估计回归参数和节点位置并检验节点效应存在性。**结果**　抑郁程度和睡眠时间存在显著的非线性关系,7~8 h 睡眠时间的抑郁程度最低,抑郁程度高分位数受睡眠时间的影响较低分位数更大。**结论**　多节点分位数回归模型拟合效果好,适用性广泛,基于模型分析结果可以为高风险人群采取更有针对性的临床和公共卫生的干预提供建议。

【关键词】 多节点分位数回归模型　抑郁　睡眠时间

【中图分类号】 R195.1　　【文献标识码】 A　　**DOI** 10.11783/j.issn.1002-3674.2024.05.003

The Application of Multi-kink Quantile Regression Model in the Study of Sleep Duration and Depression

Pan Lulu, Qin Guoyou (*Department of Biostatistics*, *School of Public Health*, *Fudan University*, *Shanghai* 200032)

【**Abstract**】 **Objective**　To introduce the Multi-kink quantile regression model and its application in the association analysis of sleep duration and depression. **Methods**　We fit the Multi-kink Quantile Regression Model to estimate regression coefficients and the kink points locations, and test the existence of the kink effect through the MultiKink package in the R software. **Results**　We identify a significant nonlinear relationship between depression and sleep duration, and the score of depression was the lowest in the sleep duration of 7~8 hours. Sleep duration has a greater impact on the higher quantile for depression. **Conclusion**　Multi-kink quantile regression model has good fitting effect and wide applicability, and the analysis results can provide more targeted advice of clinical and public health intervention for high-risk population.

【**Key words**】 Multi-kink quantile regression model; Depression; Sleep duration

在回归分析时,数据常会出现异常值、厚尾分布和模型残差项不服从同方差、正态分布的假定等情况,相比均值回归,分位数回归拟合因变量的条件分位数和自变量之间的线性关系,能全面的描述因变量条件分布的特征,不考虑同方差、正态分布的假定,对异常值更稳健[1-3]。此外,自变量和因变量之间常存在非线性关系,需要估计自变量对因变量效应变化的趋势以及效应变化的节点。样条函数在处理连续型自变量和因变量之间的非线性关系上具有优势,当自变量取值的样本点之间的间隔趋于无穷小,样本量趋于无穷大时,使用样条函数能逼近真实的曲线。但是当自变量只在有限个固定点取值时,使用样条函数将有限样本点得到的结论推广至自变量未观测到的取值范围的做法缺乏合理性[4-5]。而且在某些情况下,我们可能对效应显著改变的节点更感兴趣,而样条函数会使用曲线平滑的方式过渡多个状态,我们只能得到效应变化的大致区间。因此,使用多个节点的分段线性函数处理以上情况更为合理,既能基于有限样本点的数据拟合变量间的非线性关系,又能寻找多个效应显著的节点。

本研究介绍多节点分位数回归模型(multi-kink quantile regression model, MKQR 模型)[6],该模型将分位数回归与分段线性回归相结合,通过增加多个节点来实现自变量在不同取值范围内对因变量施加不同的影响。本研究还将该模型应用于实例数据的分析,拟合抑郁程度和睡眠时间之间的非线性关系,确定并估计节点位置;探讨在抑郁程度的不同分位数下,睡眠时间的长短对抑郁程度的影响是否一致,进一步阐明在特定情况下使用多节点分位数回归模型的优势。

原理和方法

1. 基本思想和模型概述

MKQR 模型在因变量的不同分位数下拟合不同的分段线性回归函数来表示自变量和因变量之间的非线性关系。假设 Y_t 是感兴趣的因变量,X_t 是具有阈值效应的自变量,$X_t \in [M_1, M_2]$,Z_t 表示 p 维协变量,$t = 1,2,\cdots,n$。对于给定分位数 $\tau \in (0,1)$,拟合 K 个节点的 MKQR 模型表示为

$$Q_Y(\tau | X_t, Z_t) = \alpha_0 + \alpha_1 X_t + \sum_{k=1}^{K} \beta_k (X_t - \delta_k) I(X_t > \delta_k) + \gamma^T Z_t \quad (1)$$

其中 $Q_Y(\tau | X_t, Z_t)$ 表示给定 X_t 和 Z_t 下 Y_t 的第 τ 个条件分位数,不同于均值回归中因变量的条件均值 $E(Y_t | X_t, Z_t)$。X_t 的取值范围 $[M_1, M_2]$ 被分为 $K+1$ 段,$M_1 < \delta_1 < \cdots < \delta_k < \cdots < \delta_K < M_2$,$\delta_k$ 是第 k 个节点位置,表示自变量对因变量的效应在该点发生了改变,β_k 表示 X_t 相邻两个取值范围内的斜率之差,如果 $X_t > \delta_k$,则示性函

* 基金项目:国家自然科学基金项目(82173612)

△通信作者:秦国友,E-mail:gyqin@ fudan.edu.cn

数 $I(X_t>\delta_k)=1$，若 $\beta_k\neq0$，表明相比于 X_t 的第 k 段范围，第 $k+1$ 段范围内 X_t 对 Y_t 的影响发生改变。γ 为协变量 Z_t 的系数向量。函数的斜率在节点 δ_k 处不连续，但是回归函数在自变量 X_t 的取值范围 $[M_1,M_2]$ 内都是连续的。

2. 参数估计和分析步骤

我们引用 Zhong 等[6]提出的方法和开发的工具来解决以下参数估计和假设检验问题。给定分位数 τ 和节点数 K，未知参数 $\theta=(\eta^T,\delta^T)^T$，$\eta=(\alpha_0,\alpha_1,\beta^T,\gamma^T)^T$，模型的损失函数为

$$S_n(\theta)=n^{-1}\sum_{t=1}^{n}\rho_\tau\{Y_t-Q_Y(\tau;\theta|X_t,Z_t)\},\quad \rho_\tau(u)=u\{\tau-I(u<0)\} \tag{2}$$

使用迭代分段分位数回归算法（bootstrap restarting iterative segmented quantile algorithm，BRISQ）估计回归系数和节点位置，该方法计算效率高且不受初始值影响，

$$\widehat{\theta}=(\widehat{\eta}^T,\widehat{\delta T})T=\underset{\eta\in\mathcal{B},\delta\in T}{\arg\min}S_n(\theta) \tag{3}$$

参数估计量 $\widehat{\theta}$ 被证明具有渐近正态性。

将节点数 K 的选择问题转化为模型筛选问题，基于强化分位数贝叶斯信息准则（strengthened quantile Bayesian information criterion，sBIC）选择 *sBIC* 值最小的模型，

$$sBIC(K)=\log\{S_n(\widehat{\theta_K})\}+N_k\frac{\log(n)}{2n}C_n \tag{4}$$

选出的最优模型损失函数较小且模型简单，$\widehat{K}$ 被证明具有一致性。

对于上述估计出来的节点 $\widehat{\delta}$，使用分位数得分检验法对其效应存在性进行检验。零假设（H_0）是所有节点的效应系数 β_k 均为 0，此时模型是普通的线性分位数回归；备择假设（H_1）是至少有一处节点的效应系数 β_k 不为 0。统计量为

$$T_n(\tau)=\sup_{\delta\in T}|R_n(\delta)| \tag{5}$$

由于在 H_0 下，$T_n(\tau)$ 不具有标准的渐近分布，因此采用 bootstrap 法获得近似 *P* 值进行推断。上述参数估计、模型选择和假设检验的过程均可以通过 R 包 MultiKink 来实现。

实例分析

1. 数据来源

数据来源于 2015 年至 2020 年 3 月美国全国健康和营养调查（national health and nutrition examination survey，NHANES）公开数据库收集的研究数据（https://www.cdc.gov/nchs/nhanes/index.htm）。这是一项基于人群的横断面调查，旨在收集美国成人和儿童的健康和营养状况的信息。

NHANES 采用患者健康问卷（patient health questionnaire-9，PHQ-9）评估调查对象的抑郁程度，问卷一共包括 9 个询问过去 2 周内抑郁症状出现频率的问题，回答分为“完全没有”、“几天”、“半天以上”和“几乎每天”四个类别，分值为 0~3 分，总分为 0~27 分，总分≥10 分作为临床上确定抑郁症的参考值。睡眠时间由调查对象自我回忆并报告，定义为工作日晚上平均睡眠时间，少于 3 h 和超过 14 h 被重新编码为“≤3”和“≥14”，其余睡眠时间四舍五入到最接近的半小时。基于以往研究结果[7-10]，协变量采用有向无环图来确定用于估计睡眠时间对抑郁程度影响的最小充分调整集，调整年龄、性别、种族、教育程度、家庭收入贫困比、婚姻状况、工作和体育活动和乙醇饮用状况。家庭收入贫困比是家庭收入与贫困线的比值；工作和体育活动是根据活动类型和强度计算的每周代谢当量（metabolic equivalent，MET）分数。

本研究纳入年龄范围在 20~79 岁的研究对象，排除睡眠时间、抑郁程度得分和重要协变量有缺失数据的研究对象，使用完整数据集进行分析。

2.统计分析

构建 MKQR 模型拟合抑郁程度和睡眠时间之间的非线性关联，并校正协变量：

$$Q_Y(\tau|X_t,Z_t)=\alpha_0+\alpha_1X_t+\sum_{k=1}^{K}\beta_k(X_t-\delta_k)I(X_t>\delta_k)+\gamma^TZ_t \tag{6}$$

其中 Y_t、X_t 和 Z_t 分别为抑郁程度得分、睡眠时间和协变量；$Q_Y(\tau|X_t,Z_t)$ 即 $inf\{y:F_{Y|X,Z}(y)\geq\tau\}$，表示给定睡眠时间 X_t 和协变量 Z_t 下，Y_t 条件概率分布的第 τ 个分位数（百分位数）；β_k 反映节点 δ_k 后睡眠时间和抑郁程度拟合的线段斜率的变化。我们使用 MultiKink 包中的 mkqr.bea() 函数构建 MKQR 模型，估计回归参数和节点位置，并基于 sBIC 准则进行模型筛选；使用 kinkTest() 函数进行分位数得分检验，对节点效应的存在性进行推断。本研究使用 R 4.0.3 软件进行数据分析，$P\leq0.05$ 认为差异具有统计学意义。

结　果

最终分析共纳入 6119 名研究对象，年龄中位数为 46 岁（四分位距：32~60 岁），男性占 44%（$n=2697$），以非西班牙裔白人为主，占总人数 35%（$n=2170$）。以小时为单位，将睡眠时间分为 12 个组，绘制了每个组抑郁程度得分的分布图，如图 1 所示。每组的抑郁程度得分分布均呈现典型的偏峰、右拖尾特征。

在 0.1、0.3、0.5、0.7 和 0.9 分位数水平下对抑郁程度得分和睡眠时间的关系拟合 MKQR 模型，并对节点效应进行分位数得分检验。对总人群拟合的模型参数如表 1 所示。除了 0.1 分位数外，MKQR 模型都估计了一个效应显著的节点（$P_{score}<0.001$），节点位置都在 7.5~8 h 之间。在节点位置前后，拟合的分段线性函

数都呈现先下降后上升的趋势(图 2A),表明随着睡眠时间的减少或增加,抑郁程度都会增加。然而在节点前,随着睡眠时间增加,较高分位数相比较低分位数抑郁程度得分下降更快,在节点后,随着睡眠时间增加,较高分位数相比较低分位数抑郁程度得分上升更快,且节点前后,抑郁程度的下降率和上升率在分位数水平上存在显著差异(P_{tukey}<0.001),表明在抑郁程度分位数更高的人群中,睡眠时间对抑郁程度的影响更大。

为了研究睡眠时间对抑郁程度的影响是否与性别、年龄有关,我们进行了分层分析,对不同性别(男、女)和年龄组(20~44 岁、45~59 岁、60~79 岁)的人群分别拟合了 MKQR 模型。分性别的拟合结果如图 2B~C 所示,女性相比于男性,分位数更高的人群抑郁程度受睡眠时间影响变化更大。分年龄组的拟合结果如图 2D~F 所示,年龄更大的人,抑郁程度得分相对更高,从回归线斜率的变化上看,在分位数更高的老年人中,睡眠时间更长对抑郁程度的影响可能比睡眠时间短更大。估计的睡眠时间均在 7~8 h 左右。

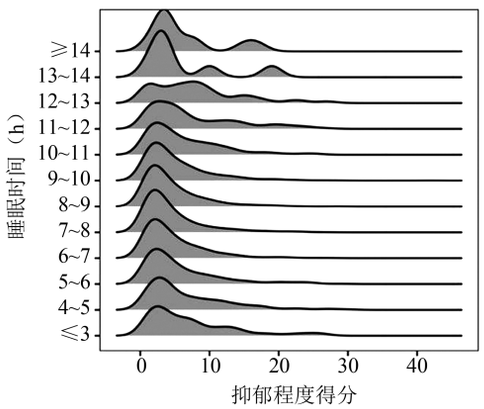
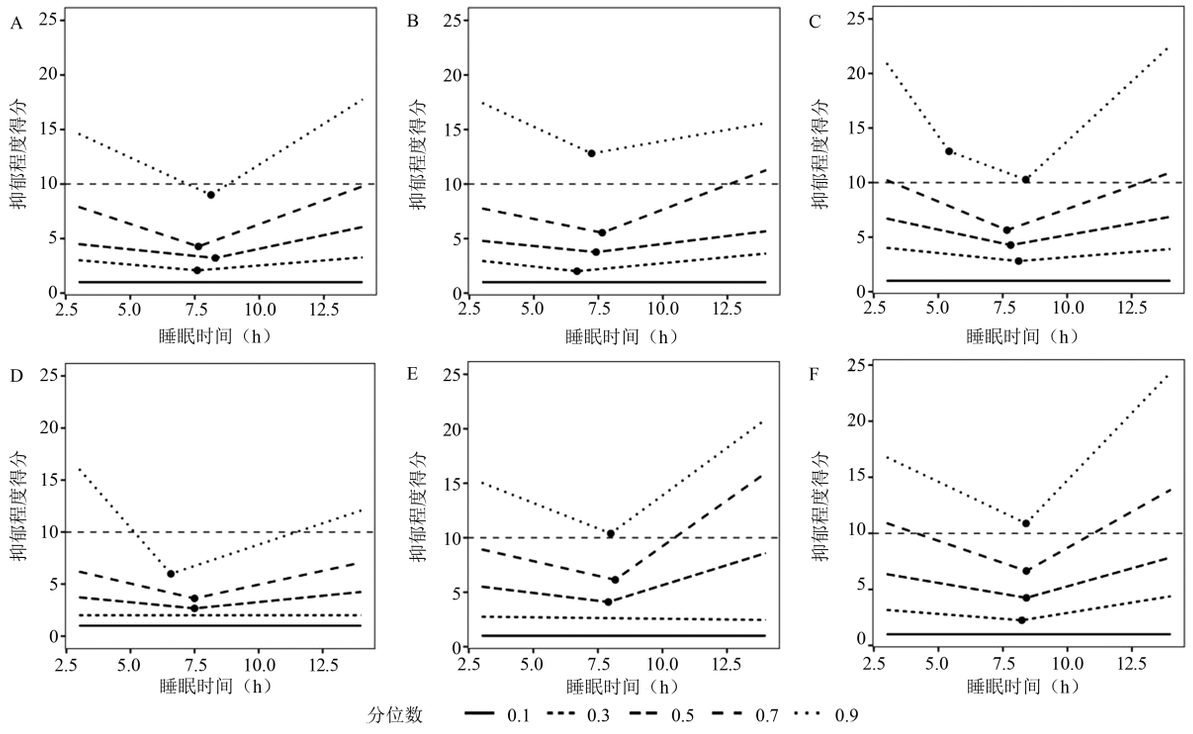


图 1　不同睡眠时间抑郁程度得分的分布图

表 1　总人群抑郁程度得分不同分位数下 MKQR 模型的拟合结果

τ	$\widehat{K}$	$\widehat{\delta_1}(SE)$	$\widehat{\alpha_0}(SE)$	$\widehat{\alpha_1}(SE)$	$\widehat{\beta_1}(SE)$	P_{score}
0.1	0	-	1	0	-	-
0.3	1	7.597(0.272)	3.629(0.424)	−0.203(0.047)	0.389(0.082)	<0.0001
0.5	1	8.294(0.251)	5.217(0.658)	−0.241(0.070)	0.740(0.126)	<0.0001
0.7	1	7.641(0.137)	10.254(1.187)	−0.783(0.143)	1.649(0.226)	<0.0001
0.9	1	8.128(0.260)	17.891(1.814)	−1.094(0.162)	2.584(0.531)	<0.0001
				P_{tukey}<0.0001	P_{tukey}<0.0001	

τ:分位数;$\widehat{K}$:估计的节点个数;$\widehat{\delta_1}$:估计的第一个节点位置;$\widehat{\alpha_0}$:截距;$\widehat{\alpha_1}$:$\widehat{\delta_1}$前回归线斜率;$\widehat{\beta_1}$:$\widehat{\delta_1}$后回归线斜率与$\widehat{\alpha_1}$之差;SE:估计量的标准差;P_{score}:对节点存在效应进行分位数得分检验的 P 值;P_{tukey}:对斜率进行 Tukey 检验并获得校正 P 值。



A:全体人群;B:男性群体;C:女性群体;D:年龄范围在 20~44 岁的人群;E:年龄范围在 45~59 岁的人群;F:年龄范围在 60~79 岁的人群;图中抑郁程度得分=10 处的水平虚线表示临床上诊断抑郁症的参考分数。

图 2　不同抑郁程度得分分位数下睡眠时间和抑郁程度拟合结果

讨 论

本研究介绍了 MKQR 模型并将其应用于抑郁程度和睡眠时间的关联分析,补充了睡眠时间和抑郁程度之间存在关联的证据,为降低抑郁风险的措施提供一些参考。研究结果显示抑郁程度和睡眠时间存在显著的非线性关系,睡眠时间过短或过长对应的抑郁程度都会升高,睡眠时间在 7~8 h 左右抑郁程度最低,该结果与以往研究得到的结果一致[7,9-10];更重要的是,本研究发现在低分位数下,睡眠时间对抑郁程度的影响很小,而在高分位数下,睡眠时间对抑郁程度的影响较大;此外高分位数下女性相比男性群体,老年人相比年轻群体,睡眠时间对抑郁程度的影响更大。因此,睡眠时间可能是抑郁症患者一个很有前景的干预目标,充足的睡眠时间(7~8 h)也许可以让抑郁症患者从中受益。

对于抑郁程度和睡眠时间的关联分析及类似研究,国内外许多研究者采用样条函数来拟合自变量和因变量之间的非线性关系[7,11]。样条函数本质上是一组平滑连接的分段多项式函数,被广泛用于拟合连续型自变量和因变量之间的非线性关系[5,12],但是有的情况下自变量并不连续,例如吸烟量(<1 包/周、2~3 包/周、3~5 包/周,>5 包/周)、教育水平(未接受教育、小学毕业、初中毕业、高中毕业、本科毕业、硕士毕业、博士毕业)和抑郁评分(0~27 分,1 分为 1 单位)。对于以上情况,我们只能基于自变量的固定取值点进行分析,无法对自变量进一步细分,即使扩大样本量,我们也无法确定使用样条函数拟合的曲线能否将有限样本点的结论推广至自变量未观测的取值范围[4-5]。因此,针对以上资料建立 MKQR 模型推断有限样本点之间效应的变化,相比使用样条函数估计平滑曲线更为合理,且具有较好的拟合效果[13]。

此外,有的情况下我们对效应显著的节点位置和节点个数更感兴趣。样条函数需要人为指定节点位置和个数,但是一般情况下,没有足够的专业背景知识指导自变量和因变量的关系在哪些特定的位置转折,因此样条函数的节点选择带有主观性[14-15],而节点的个数会影响曲线的形状以及平滑程度,当样条设定的节点越多,模型越复杂,结果也越难以解释,且样条函数得到的平滑曲线估计只能得到效应变化的大致区间。使用 MKQR 模型不仅能基于数据自由推断多个节点的位置和个数,检验节点的效应是否显著,而且需要估计的参数相对较少,模型相对简单。

综上所述,MKQR 模型适用于因变量不服从正态分布,自变量取值离散、有限的情况,能探索变量间的非线性关系,寻找效应存在的节点位置,对因变量的条件分布进行全面描述。MKQR 模型灵活性强,适用范围广,可以应用于相关流行病学和医学研究。从公共卫生的角度看,MKQR 模型得到的结论丰富,可以得到感兴趣人群的分析结果,可以为高风险人群采取更有针对性的临床和公共卫生的干预提供建议。

参 考 文 献

[1] 郭月玲, 李春波. 分位数回归理论及其应用[J]. 吉首大学学报(自然科学版),2014,35(5):26-28.

[2] 吴建南, 马伟. 估计极端行为模型:分位数回归方法及其实现与应用[J]. 数理统计与管理,2006,25(5):536-543.

[3] 赵为华. 分位数回归在比例数据分析中的应用[J]. 南通大学学报(自然科学版),2016,15(1):71-76.

[4] Racine J, Li Q. Nonparametric estimation of regression functions with both categorical and continuous data[J]. Journal of Econometrics,2004,119(1):99-130.

[5] Ma S, Racine J S, Yang L. Spline regression in the presence of categorical predictors[J]. Journal of Applied Econometrics, 2015, 30(5):705-717.

[6] Zhong W, Wan C, Zhang W. Estimation and inference for multi-kink quantile regression[J]. Journal of Business & Economic Statistics,2022, 40(3):1123-1139.

[7] Dong L, Xie Y, Zou X. Association between sleep duration and depression in US adults: A cross-sectional study[J]. J Affect Disord, 2022,296:183-188.

[8] Iranpour S, Sabour S. Inverse association between caffeine intake and depressive symptoms in US adults: data from National Health and Nutrition Examination Survey(NHANES)2005—2006[J]. Psychiatry Res,2019,271:732-739.

[9] Ji S, Wang J, Wang W, et al. Longer depressive duration reduces sleep duration more: A longitudinal study in the middle-aged and elderly Chinese[J]. J Affect Disord,2022,317:185-192.

[10] Liao F, Wang W, Zhou B, et al. Longitudinal Cohort Study of the Relationship between Sleep Duration and Depressive Symptoms in Older People in China[J]. Journal of Sichuan University(Medical Sciences),2022, 53(1):109-113.

[11] Smiley A, King D, Bidulescu A. The Association between Sleep Duration and Metabolic Syndrome: The NHANES 2013/2014[J]. Nutrients,2019,11(11):2582.

[12] Ma S. Theory of Spline Regression with Applications to Time Series, Longitudinal, and Categorical Data, and Data with Jumps[M]. Michigan State University. Statistics,2011.

[13] Varol B, Omurlu I M K, Mevlüt T. Comparison of piecewise regression and polynomial regression analyses in health and simulation data sets[J]. Süleyman Demirel Üniversitesi Sağlık Bilimleri Dergisi, 2020,11(2):144-151.

[14] 王晓晓, 陶立元, 李楠, 等. 限制性立方样条在非线性关联分析中的应用[J]. 中华儿科杂志,2020, 58(8):652.

[15] 罗剑锋, 金欢, 李宝月,等. 限制性立方样条在非线性回归中的应用研究[J]. 中国卫生统计,2010,27(3):229-232.

(责任编辑:邓 妍)