基于线性混合模型树在体质指数纵向轨迹中的应用

东南大学公共卫生学院流行病与卫生统计系(210009) 臧一腾 陈思臻 陆贝尔 缪鹏程 马溶基 陈炳为△

【提 要】 **目的** 探讨江苏省成年人体质指数纵向变化轨迹及分类。**方法** 基于中国营养健康调查数据,使用线性混合模型树探究江苏省 18~65 岁人群体质指数的变化轨迹及其分类情况。**结果** 线性混合模型树生成了 13 个节点,树的深度为 6,分类节点为基线 BMI、平均摄入卡路里、基线年龄。**结论** 线性混合模型树可以识别体质指数的变化轨迹,拓展了纵向数据动态变化的研究方法。

【关键词】 身体质量指数 线性混合模型树 纵向数据
【中图分类号】 R195.1 【文献标识码】 A **DOI** 10.11783/j.issn.1002-3674.2024.01.008

Application of Linear Mixed Model Tree in Longitudinal Trajectory of Body Mass Index

Zang Yiteng, Chen Sizhen, Lu Beier, et al (*Department of Epidemiology and Health Statistics, School of Public Health, Southeast University* (210009), *Nanjing*)

【**Abstract**】 **Objective** To understand the trajectory and classification of adult body mass index (BMI) in Jiangsu Province. **Methods** Based on China Health and Nutrition Survey, this study used the linear mixed model tree to explore the trajectory and classification of BMI of people aged 18-65 in Jiangsu Province. **Results** The linear mixed model tree had 13 nodes and the depth was 6. The classification nodes were baseline BMI, average calorie intake and baseline age. **Conclusion** The linear mixed model tree can identify the trajectory of BMI and expand the research method of longitudinal data.

【**Key words**】 Body mass index; Linear mixed model trees; Longitudinal data

肥胖是全球公共卫生的一个重要问题,中国是世界上肥胖人口最多的国家之一,肥胖已成为我国卫生保健系统面临的重大挑战[1]。已有的超重肥胖趋势分析发现,2002 年我国 18 岁以上成人超重率为 22.8%,肥胖率为 7.1%;2006 年 18 岁以上成人超重率为 28.8%,肥胖率为 8.4%;2015 年超重率为 34.8%,肥胖率为 14.5%[2];中国各地区调查也发现超重和肥胖发生率处于增长的趋势。近几十年来,纵向数据的统计方法不断发展,最常用的方法有重复测量方差分析、广义估计方程和混合效应模型等。本文以江苏省 2002-2015 年的成年人为例,利用线性混合模型树(linear mixed model trees, LMM Trees)对身体质量指数(body mass index, BMI)随时间的变化轨迹进行分析。

资料与方法

1.数据来源

中国营养健康调查(China health and nutrition survey, CHNS)是中国疾病预防控制中心营养与食品安全所与美国北卡罗莱纳大学人口中心合作,从 1989 年起面向中国 9 省开展的居民膳食结构与营养状况变迁的追踪研究。在 1989-2015 年期间对同一人群的社会经济状况、卫生服务、居民膳食结构和营养状况进行了追踪调查。本文选取江苏省 2002-2015 年,至少参与 3 次调查的人群,研究 15 年间身体质量指数变化轨迹情况。纳入人群年龄 18~65 岁,排除当年怀孕者、糖尿病、肿瘤患者以及极端 BMI 情况,最终纳入 908 人。协变量包括基线年龄、性别、饮酒与吸烟史、体力劳动强度、基线 BMI、平均摄入卡路里、蛋白质供能比例,碳水化合物供能比例,基线定义为第一次接受调查的时间。

2.研究方法

在决策树算法中,使用递归方法能够选择最优特征(预测变量),并根据所选特征对训练数据进行分割,形成较好的分类。决策树模型优点在于能够自动检测特征间可能存在的交互作用,通过图形易于对模型的解释。在递归分割的思想基础上,参数模型被集成到树中,将递归分割嵌入到统计模型估计和变量选择中,称为基于模型的递归分割(model-based recursive partitioning, MOB)。在此框架内,通过计算分支树来拟合分段参数模型,其中每片叶子都与拟合模型相关联[3]。MOB 基本思想是每个节点都是基于某种类型的统计模型(如线性回归、线性混合模型等),目标函数用于估计参数和分割点。递归分割可以对非线性关系进行建模,并自动检测解释变量之间的交互作用。该算法使得数据信息的层次结构得以体现,模型结果的解释性更佳[4]。

Fokkema 等[5]最初提出了广义线性混合模型树(generalized linear mixed model trees, GLMMT 或

△通信作者:陈炳为,E-mail:drchenbw@126.com

glmertrees)，它是 MOB 模型中的一种，树中的节点由特定子数据的广义线性混合模型组成，允许在广义线性混合效应模型中检测子数据相互作用和非线性的模型。GLMMT 由全局和局部两部分组成：全局模型由随机效应项和所有观测值组成。局部模型由局部估计的固定效应项组成：数据集中的观测值根据附加协变量进行划分，并在所得划分的每个单元中估计单独的固定效应模型。

对于 i 簇子数据，协变量 χ_i，应变量 Y_i，广义线性混合效应模型则由以下公式得出：

$$g(\mu_{ij})=\chi_i^T\beta_j+Z_i^Tb$$

g 是链接函数，固定效应 β_j 是局部参数，其值取决于终端节点 j，随机效应 b 是全局的。公式表示只有随机截距的混合效应模型，Z_i 是常数，b 是与簇 i 相关的随机截距[5-6]。GLMMT 的步骤如下：

（1）将 r 和所有值 $\hat{b}_{(r)}$ 初始化为 0；

（2）设置 $r=r+1$，使用 $z_i^T\hat{b}_{(r-1)}$ 估计广义线性模型树；

（3）使用步骤（2）中估计的树中的终节点 $j(r)$ 拟合混合效应模型 $g(\mu_{ij})=\chi_i{}^T\beta_j+z_i{}^Tb$。从估计模型中提取后验预测 $\hat{b}_{(r)}$；

（4）重复（2）和（3），直到收敛。

在每次迭代中，在步骤（2）中重新估计树，在步骤（3）中重新估计固定和随机效应参数。在 GLMMT 中，应变量可以是连续、二分类或计数变量；特征可以是连续或分类变量。当链接函数 g 为恒等函数时，GLMMT 即为 LMM Trees。GLMMT 在模型预测中显示出了一定的优势，Fokkema 等研究者[7]应用 GLMMT 研究接受心理健康服务的 3256 名年轻人的治疗效果，同时考虑 18 个人口学、严重程度等特征，发现 GLMMT 预测准确性与传统 GLMM 和随机森林的预测准确性基本相同，但 GLMMT 模型中纳入的变量更少。

本文中连续响应变量为调查年份的 BMI，线性模型的预测变量为年份，潜在的分割变量有基线年龄、性别、饮酒与吸烟史、体力劳动强度、基线 BMI、平均摄入卡路里、蛋白质供能比例、碳水化合物供能比例，同时考虑个体的随机效应影响。统计分析使用 R 4.0.5 的“glmertree”包，该包使用“partykit”包确定分区，使用“lme4”包估算混合效应模型。由于结局变量为连续型变量，本研究调用“glmertree”包中线性混合模型树“lmertree”功能进行统计分析。$P<0.05$ 被认为具有统计学差异。

结　果

1.人群基本特征

经过筛选和匹配，最终纳入符合条件者 908 人，共 3884 条调查记录，其中男性 420 人，占比 46.26%；基线年龄为 40.40±11.26 岁。其他潜在分割变量特征见表 1。

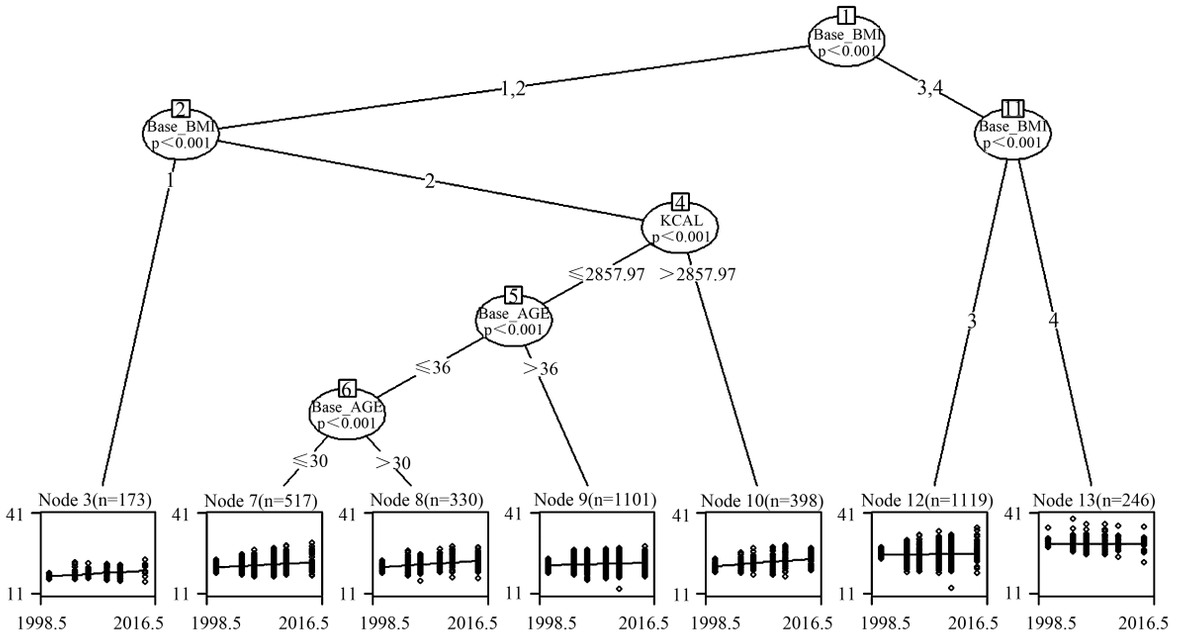
表 1　纳入人群基本特征及潜在分割变量

变量	人数（构成比，%）/均数±标准差
性别	
男	420（46.26）
女	488（53.74）
基线 BMI（kg/m^2）	
<18.5	46（5.07）
18.5~	545（60.02）
24~	258（28.41）
≥28	59（6.50）
是否吸烟	
是	336（37.00）
否	572（63.00）
是否饮酒	
是	495（54.52）
否	413（45.48）
体力劳动强度*	
极轻体力活动	196（21.59）
轻体力活动	232（25.55）
中度体力活动	174（19.16）
重体力活动	306（33.70）
基线年龄（岁）	40.40±11.26
平均摄入卡路里（千卡）	2399.93±536.76
平均蛋白质供能比例	13.32±1.91
平均碳水化合物供能比例	60.00±8.78

*：极轻体力活动：坐着工作、如办公室工作人员、修表工等；轻体力活动：站着工作、如售货员、实验室技术人员、教师等；中度体力活动：学生、司机、电工、金属制造工人等；重体力活动：农民、舞蹈演员、钢铁工人、运动员等。

2.LMM Trees 分析结果

线性混合效应模型树以 $P=0.05$ 为界值，树的深度为 6，共形成 13 个节点，如图 1。基线 BMI、平均摄入卡路里、基线年龄为子节点，共分成了 7 类变化轨迹。最终拟合的模型体现了剩余（固定-随机效应）预测变量均值的变化。根据表 2 中各个终结点的估计系数可知，随着时间的变化，第一次调查偏瘦的人群 BMI 呈现上升趋势，斜率为 0.142；第一次调查 BMI 正常且摄入卡路里小于 2857.97 千卡/天的人群中，年龄小于等于 30 岁的人群，BMI 上升斜率为 0.116，年龄大于 30 岁且小于 36 岁的人群，BMI 上升斜率为 0.158，年龄大于 36 岁的人群，BMI 上升斜率为 0.062；第一次调查 BMI 正常且摄入卡路里大于 2857.971 卡/天的人群，BMI 上升斜率为 0.181；第一次调查超重的人群，BMI 上升斜率为 0.037；第一次调查肥胖的人群，BMI 呈下降趋势，变化斜率为-0.007。模型组内相关系数（intraclass correlation coefficient，ICC）为 0.6298。图 2 显示了线性混合效应模型树每个终节点系数及其 95% *CI*。模型平均绝对百分比误差（mean absolute percentage error，MAPE）为 3.576%。将整体以患者为单位平均分为 10 份，进行十折交叉验证，结果显示 MAPE 为7.094%，结果表明模型的拟合较为理想。



*：Base_BMI：基线体质指数（kg/m²），BMI-1：基线体质指数<18.5kg/m²，BMI-2：基线体质指数 18.5～24 kg/m²，BMI-3：基线体质指数 24～28kg/m²，BMI-4：基线体质指数≥28kg/m²；KCAL：平均摄入卡路里（千卡）；Base_AGE：基线年龄（岁）。

图 1 混合效应模型树

表 2 线性混合效应模型树在终节点上的估计系数

终结点	截距	斜率
Node 3	−267.084	0.142
Node 7	−212.035	0.116
Node 8	−294.488	0.158
Node 9	−101.878	0.062
Node 10	−340.667	0.181
Node 12	−48.341	0.037
Node 13	44.193	−0.007

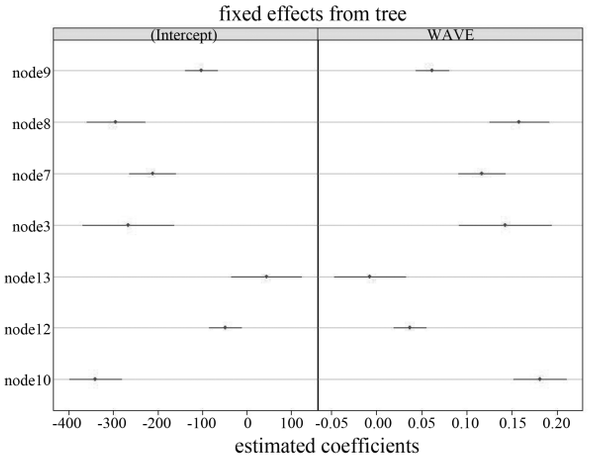


图 2 线性混合效应模型树终节点的系数估计图

讨　论

GLMMT 常用于以下 3 种情况：聚类（多级）数据的终节点上常数拟合、治疗亚组相互作用的检测、纵向生长曲线模型中的亚组检测[5]。广义线性混合模型在线性预测的部分引入随机效应，解释重复测量结果的相关性[8]。除了寻找结果的预测因子外，广义线性混合模型树被认为在处理纵向数据（重复测量资料）时具有独特的优势，还可以用于多层次结构的数据[7]，包括来自多个试验中心的观测数据[5]，在此类数据集中，个体观察值嵌套在更高级别的单元中。

本文综合考虑了研究人群的基本特征、营养与运动情况和生活习惯，探讨影响 BMI 变化轨迹的潜在分类变量。本研究发现，15 年来，调查人群的 BMI 呈上升趋势。在线性混合效应模型树中，基于模型的递归划分思想将 18～65 岁人群 BMI 变化轨迹划分成 7 类，类别受到第一次调查时 BMI 水平、平均摄入卡路里、第一次调查年龄影响。本研究使用 MAPE 作为模型预测准确性的评价指标，十折交叉验证结果表明，LMM Trees 具有较好的预测性能。

在本研究中，基线 BMI 作为第一个分割变量，说明基线 BMI 是体重变化中的重要影响因素。BMI 作为身高和体重的综合指标，与基础代谢率关系紧密[9]。成年人基线 BMI 与遗传易感程度具有交互作用[10]，影响 BMI 的变化轨迹。基线 BMI ≥ 28kg/m² 是唯一一个呈现 BMI 下降趋势的分区，可能的原因是随着人们健康意识的提高，肥胖人群有意识地控制自己的体重、调节饮食和生活方式。Hayes 等发现[11]，从 1980 年代到现在，BMI 年度增幅最大的改善（降低）发生在 BMI 最高百分位数的年轻女性中。超重人群的 BMI 处于上升趋势，主要原因可能是超重人群 BMI 尚未达到肥胖标准，生活未受到影响，未能引起重视。夏云婷等[12]对 2013 年中国 18 岁及以上高血压患者进行了体重自评的研究，发现超重人群体重自评准确率仅为 24.9%，不利于控制体重。第一次调查偏瘦的人群 BMI 呈现上升趋势，这是一个较为健康的

转变,但整体依然处于较低水平。

《中国居民膳食指南(2016 版)》建议,男性每日建议摄入热量为 2250 千卡,女性为 1800 千卡。在本研究中,第一次调查处于正常的人群,以平均摄入卡路里 2857.97 千卡/天为分割点,摄入热量大于 2857.97 千卡/天的人群,BMI 上升速度较快。

已有的研究表明,BMI 的变化轨迹在不同的年龄组表现不同[11,13]。我们的研究发现,对于第一次调查 BMI 处于正常范围的人群,36 岁前 BMI 上升趋势较为明显。相似的,在 Fang 等人的研究中[14],年轻人群(尤其是男性)的 BMI 增长快于年长人群,这一发现可能与中国经济发展和国内生产总值从低到中高的转变有关,影响了人们的一般生活方式。现有关于 BMI 变化趋势的研究大多是关注肥胖、超重发病率的总体变化趋势,很少有研究对于调查人群进行特征分类,以得到具有共性的变化轨迹。本研究介绍了基于模型的递归分割方法及应用,从分类的角度提出了探索 BMI 变化轨迹的新思路,广义混合效应模型树的应用更将推进精准医疗的发展。但值得注意的是 BMI 的长期变化受到环境和个人因素的影响,更多社会心理因素对于 BMI 的影响有待于纳入考虑。另一方面,BMI 本身也存在一定的局限性。BMI 不能很好地预测体成分,不能完全体现个人体脂率等指标[15],对于肥胖的判断存在一定偏差。

参 考 文 献

[1] Zeng Q, Li NS, Pan XF, et al. Clinical management and treatment of obesity in China. The Lancet Diabetes & Endocrinology, 2021, 9(6): 393-405.

[2] 朱莹, 徐宁. 2006-2015 年我国成人超重和肥胖长期变化趋势分析. 中国药物与临床, 2020, 20(11): 1803-1804.

[3] Zeileis A, Hothorn T, Hornik K. Model-Based Recursive Partitioning. Journal of Computational and Graphical Statistics, 2008, 17(2): 492-514.

[4] Pellagatti M, Masci C, Ieva F, et al. Generalized mixed effects random forest: A flexible approach to predict university student dropout. Statistical Analysis and Data Mining, 2021, 14(3): 241-257.

[5] Fokkema M, Smits N, Zeileis A, et al. Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. Behavior Research Methods, 2018, 50: 2016-2034.

[6] Xu Y, Zafirov A, Alvarez RM, et al. FREEtree: A Tree-based Approach for High Dimensional Longitudinal Data With Correlated Features. arxiv.org/abs/2006.09693, 2020.

[7] Fokkema M, Edbrooke-Childs J, Wolpert M. Generalized linear mixed-model (GLMM) trees: A flexible decision-tree method for multilevel and longitudinal data. Psychotherapy Research, 2021, 31(3): 329-341.

[8] 汤宁, 宋秋月, 易东, 等. 医学纵向数据建模方法及其统计分析策略. 中国卫生统计, 2019, 36(3): 441-444+447.

[9] 张莹, 吴景欢, 洪平, 等. 北京市超重和肥胖成人基础代谢率的研究. 卫生研究, 2016, 45(5): 739-742.

[10] Song M, Zheng Y, Qi L, et al. Longitudinal Analysis of Genetic Susceptibility and Body Mass Index throughout Adult Life. Diabetes, 2018, 67(2): 248-255.

[11] Hayes AJ, Gearon E, Backholer K, et al. Age specific changes in BMI and BMI distribution among Australian adults using cross-sectional surveys from 1980to 2008. International Journal of Obesity, 2015, 39(8): 1205-1216.

[12] 夏云婷, 刘少博, 王丽敏, 等. 2013 年我国 18 岁及以上高血压患者体重测量及自评情况. 中国健康教育, 2019, 35(8): 685-690.

[13] 高仲淳, 邹波, 蓝恭赛, 等. 20~59 岁成年人体质指数随年龄变化轨迹与高血压发病的关系研究. 中国全科医学, 2021, 24(8): 954-958.

[14] Fang CC, Liang Y. Social disparities in body mass index (BMI) trajectories among Chinese adults in 1991-2011. International Journal for Equity in Health, 2017, 16(1): 146.

[15] 张强. 体质指数和体脂肪率评价成年人肥胖的比较. 卫生研究, 2019, 48(4): 573-576.

(责任编辑:张　悦)

(上接第 40 页)

[15] Hashtarkhani S, Kiani B, Bergquist R, et al. An age-integrated approach to improve measurement of potential spatial accessibility to emergency medical services for urban areas. International Journal of Health Planning and Management, 2020, 35(3): 788-798.

[16] Chu HJ, Lin BC, Yu MR, et al. Minimizing Spatial Variability of Healthcare Spatial Accessibility-The Case of a Dengue Fever Outbreak. International Journal of Environmental Research and Public Health, 2016, 13(12): 1235-1245.

[17] Fransen K, Neutens T, De Maeyer P, et al. A commuter-based two-step floating catchment area method for measuring spatial accessibility of daycare centers. Health Place, 2015, 32: 65-73.

[18] 沈玉卿, 白灵瑶, 程杨, 等. 基于综合指标体系评价北京市二、三级医院住院服务的空间可及性. 中国卫生信息管理杂志, 2020, 17(5): 675-681.

[19] 罗力, 付晨, 吴凌放, 等. 医疗服务地理可及性及其可视化表达研究概述. 中国卫生资源, 2016, 19(4): 264-269.

[20] Yang N, Chen S, Hu W, et al. Spatial Distribution Balance Analysis of Hospitals in Wuhan. International Journal of Environmental Research and Public Health, 2016, 13(10): 971-986.

[21] Kiani B, Bagheri N, Tara A, et al. Comparing potential spatial access with self-reported travel times and cost analysis to haemodialysis facilities in North-eastern Iran. Geospatial Health, 2018, 13(2): 703-709.

[22] 徐斌, 覃青连, 韦雪, 等. 南宁市社区卫生服务的空间可及性评估: 高斯两步移动搜索法. 中国卫生统计, 2021, 38(6): 852-859.

[23] Tuson M, Yap M, Kok MR, et al. Incorporating geography into a new generalized theoretical and statistical framework addressing the modifiable areal unit problem. International Journal of Health Geographic, 2019, 18(1): 6-20.

[24] Ranga V, Panda P. Spatial access to inpatient health care in northern rural India. Geospatial Health, 2014, 8(2): 545-556.

(责任编辑:张　悦)