计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(10):152−162 [doi: 10.15888/j.cnki.csa.009658]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

数据流下部分线性模型的在线估计①

卢果林

(中国科学技术大学 人工智能与数据科学学院, 合肥 230026)
通信作者: 卢果林, E-mail: luguolin@mail.ustc.edu.cn

摘　要: 部分线性模型作为一种重要的半参数回归模型, 因其在复杂数据结构分析中表现出的灵活适应性, 广泛应用于各领域. 然而, 在大数据背景下, 该模型的研究和应用面临着多重挑战, 其中最为关键的难点在于计算速度和数据存储. 本文针对以数据块形式连续观测的数据流场景, 提出一种在线估计的计算方法, 用于估计部分线性模型中线性部分的参数和非线性部分的未知函数. 该方法仅需利用当前数据块和之前计算过的汇总统计量即可实现实时估算. 数值模拟从两个角度进行验证有效性: 分别改变数据流的单位数据块大小和总样本规模, 以比较在线估计方法和传统估计方法的偏差、标准误差以及均方误差. 实验表明, 与传统方法相比, 本文的方法具有快速计算和无需重新访问历史数据的优势, 同时在均方误差方面接近传统方法. 最后, 基于中国综合社会调查 (CGSS) 数据, 本文应用在线估计方法分析我国劳动年龄人口生活质量的影响因素, 得出周工作时间在 30–60 h 范围内的全职工作对提升生活质量具有积极作用的结论, 为相关政策制定提供了一定参考价值.

关键词: 在线估计; 部分线性模型; 核回归; 大数据; 数据压缩

引用格式: 卢果林.数据流下部分线性模型的在线估计.计算机系统应用,2024,33(10):152–162. http://www.c-s-a.org.cn/1003-3254/9658.html

Online Estimation for Partially Linear Model in Data Streams

LU Guo-Lin

(School of Artificial Intelligence and Data Science, University of Science and Technology of China, Hefei 230026, China)

Abstract: The partially linear model, as an important type of semiparametric regression models, is widely used across various fields due to its flexible adaptability in the analysis of complex data structures. However, in the era of big data, the research and application of this model are faced with multiple challenges, with the most critical ones being computing speed and data storage. This study considers the scenario of data streams continuously observed in the form of data blocks and proposes an online estimation method for the parameters of the linear part and the unknown function of the nonlinear part in the partially linear model. This method enables real-time estimation using only the current data block and previously computed summary statistics. To verify the effectiveness, the unit data block size and the total sample size of the data streams are changed respectively in numerical simulations, so that the bias, standard error and mean squared error between the online estimation method and the traditional one can be compared. The experiments demonstrate that, compared to the traditional method, the proposed approach offers the advantages of rapid computation and unnecessary review of historical data, while being close to the traditional method in terms of mean squared error. Finally, based on the data from the China general social survey (CGSS), this study applies the online estimation method to analyze the factors influencing the quality of life of the working-age population in China. The results indicate that full-time work within the range of 30 to 60 hours per week positively contributes to improving the quality of life, providing valuable references for relevant policy formulation.

① 收稿时间: 2024-01-25; 修改时间: 2024-02-29, 2024-05-06; 采用时间: 2024-05-21; csa 在线出版时间: 2024-08-21
CNKI 网络首发时间: 2024-08-22

Key words: online estimation; partially linear model; kernel regression; big data; data compression

部分线性模型兼具参数模型和非参数模型之长, 以其灵活性著称. 自从 Engle 等[1]首次提出该模型用于描述城市气候对电力消耗的影响以来, 已获得众多学者关注, 涌现出大量相关研究[2–4]. 样条方法[5]、分段多项式近似方法[6], 以及缺失数据的情况[7–9]等, 都已被纳入这一模型的讨论中. 这些工作极大丰富了部分线性模型的理论和实践应用. 然而, 它们并未考虑大数据应用的两个重大挑战: 计算时间和数据存储. 尤其在数据流场景下, 随着数据的不断积累, 存储需求激增, 每次新数据收集之后都需要对整个数据集重复进行统计分析. 冗杂的计算过程和过高的计算成本, 使得部分线性模型的传统估计计算方法变得不切实际.

近年来, 大数据模型的分析主要遵循 3 种方法. 第 1 种是子采样方法 (subsampling-based)[10–14], 该方法在回归模型中得到有效应用, 能有效处理大多数大数据场景. 然而, 经过子抽样获得的有限样本可能会遗漏关键信息, 导致估计不准确; 第 2 种方法是分治集成法[15–18] (divide-and-conquer), 其主要思路是将大数据集划分为若干子集从而把问题化繁为简. 例如, 在部分线性模型研究中, Zhao 等[19]提出的集成核岭回归方法、Lian 等[20]基于分治集成的投影样条方法. 然而分治集成法和子采样法均依赖于完整数据集, 在现实问题中, 大数据的完整收集通常需要漫长的等待时间. 如果是数据流形式的大数据, 数据更是持续生成而不会停止.

除了前面提到的两种方法, 第 3 种用于处理大数据模型的方法就是顺序更新法 (sequential-updating). 该方法特别适用于数据流, 因为它能够实时分析, 无需大量存储. 本文后面内容将这种方法称作在线估计方法, 英文写作“ONLINE”, 以区别于每次需要使用所有原始数据重新运行分析的传统方法, 把传统方法写作“OFFLINE”. 近年来关于在线估计方法的研究多见于线性回归模型[21–24], 对于非参数回归模型, Kong 等[25]提出了一种基于核回归的在线估计方法, 并将其扩展到变系数模型. 由于部分线性模型是一种特殊的变系数模型, 所以也能勉强算作一种部分线性模型的在线估计方法. 然而该方法是根据每一个新到数据点更新估计, 并且基于协变量固定在某点处进行在线计算. 如果使用这种方法处理部分线性模型, 需要假设变系数模型的某些函数项为常数以作为部分线性模型的参数 β, 这就会使得当固定点选择不当时, 核估计的边界效应将会使得这些函数项的估计效果大打折扣, 参数β的估计随着固定点的不同而有较大波动, 可能导致较大的偏差. 因此, 当前亟需一种能够有效处理数据流场景下部分线性模型的分析方法.

本文旨在解决将部分线性模型应用于大数据流时遇到的分析挑战. 为此提出了一种基于部分线性模型两步估计方法的在线更新估计方法. 与以往的研究不同, 本文方法在更新估计参数时不需要考虑选取协变量为合适的给定值. 对于非线性部分, 本文基于 Nadaraya-Watson 回归和局部线性核回归给出两个在线更新估计量. 通过对线性部分和非线性部分的分开估计从而完成了整个部分线性模型的在线估计. 本文的方法以数据块为单位进行处理, 仅利用当前数据和每次更新的一些汇总统计量, 从而确保了每次更新计算时间的基本一致. 与 OFFLINE 方法相比, 该方法不会因总样本量的增加而导致计算时间和存储需求的不断增加, 因此在计算效率和成本控制方面具有明显优势.

1 部分线性模型与其大数据估计计算问题

本节首先给出本文研究的对象, 即部分线性模型, 然后再阐述传统方法在数据流场景下所面临的计算难题.

1.1 部分线性模型

假设有 n 个 i.i.d.随机样本$\{(y_i,x_i,z_i),i=1,2,\cdots,n\}$, 部分线性模型通常表示为:

$$y_i=\boldsymbol{x}_i^{\top}\beta+g(z_i)+\varepsilon_i,\quad i=1,\cdots,n \tag{1}$$

其中, 协变量$x_i\in\mathbb{R}^p$、$z_i\in\mathbb{R}^q$, 响应变量$y_i\in\mathbb{R}$. β为有限维的参数向量, $g(\cdot)$是一个未知的光滑函数, ε_i是随机误差项.

1.2 数据流下传统估计方法的计算问题

考虑样本以数据流的形式依次独立被收集, 根据 Li[26]的参数估计方法, 有:

$$\hat{\beta}=\left[\sum_{i=1}^{n}\tilde{x}_i\tilde{x}_i^{\top}\hat{f}_i^2\right]^{-1}\sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i\hat{f}_i^2 \tag{2}$$

其中, $\tilde{x}_i = x_i - \hat{x}_i$, $\tilde{y}_i = y_i - \hat{y}_i$, $\hat{f}_i$ 是 z_i 处的核密度估计; $\hat{x}_i$ 和 $\hat{y}_i$ 分别是条件期望 $\mathbb{E}(x_i|z_i)$ 和 $\mathbb{E}(y_i|z_i)$ 的 NW 估计, 即 $\hat{x}_i = \sum_{j=1}^{n} \omega_j(z_i)x_j$, $\hat{y}_i$ 的计算同理, 这里 $\omega_j(\cdot)$ 是 NW 估计中所使用的权函数.

式 (2) 的传统方法在数据流场景下的计算问题在于, 每次新数据的到来, 使用式 (2) 进行更新计算 $\hat{\beta}$, 都需要访问所有历史数据重新计算估计量集合 $\left\{\left(\hat{x}_i, \hat{y}_i, \hat{f}_i\right), i = 1,2,\cdots,n\right\}$. 因此, 在计算时间方面, 随总样本量 n 的不断增加, 每次更新所耗时间将不断增加, 且没有上限. 在存储方面, 由于原始数据需要完整保留, 存储消耗也将持续增加. 此外, 非参数部分 $\mathbb{E}\left(y_i - x_i^{\top}\hat{\beta}|z_i\right) = g(z_i)$ 的估计, 以及所用到核估计中带宽 h 的重新计算也涉及对所有历史数据的访问, 意味着同样的问题. 上述要求使得传统方法对数据流下部分线性模型的计算难以在单台计算机上实现.

2 部分线性模型的在线更新估计方法

考虑流数据以数据块的形式依次独立被收集, 则式 (1) 的模型改写为:

$$y_l = X_l\beta + g(Z_l) + \varepsilon_l, \quad l = 1,2,\cdots \tag{3}$$

其中, 样本矩阵 $X_l \in \mathbb{R}^{n_l \times p}$、$Z_l \in \mathbb{R}^{n_l \times q}$, 向量 $y_l \in \mathbb{R}^{n_l}$. 第 l 个数据块 $\{(y_{l,m}, x_{l,m}, z_{l,m}), m = 1,2,\cdots,n_l\}$ 是样本量为 n_l 的数据子集. 记在终点 K 处的总样本量表示为 $N_K = \sum_{k=1}^{K} n_k$, 总样本矩阵表示为 $X = (X_1^{\top},\cdots,X_K^{\top})^{\top}$, $Z = (Z_1^{\top},\cdots,Z_K^{\top})^{\top}$, 响应变量 $y = (y_1^{\top},\cdots,y_K^{\top})^{\top}$.

本节将详细介绍针对式 (3) 的模型而设计的在线估计方法, 以解决第 1 节中传统方法带来的问题. 该方法基于两步估计过程, 首先是对模型参数部分的在线估计, 这一步骤涉及对模型中的参数进行动态更新. 然后是非参数部分的在线估计, 给出基于局部核回归的具体计算过程以及带宽的选取方式.

2.1 参数部分的在线更新估计

当第 k 个数据块到达时, 单独使用该样本子集对参数获得一个估计 $\hat{\beta}_{k,n_k}$, 由式 (2) 有:

$$\hat{\beta}_{k,n_k} = \left(\tilde{X}_k^{\top}W_k\tilde{X}_k\right)^{-1}\left(\tilde{X}_k^{\top}W_k\tilde{y}_k\right) \tag{4}$$

其中, $W_k = \mathrm{diag}\{\hat{f}_{k,1}^2,\cdots,\hat{f}_{k,n_k}^2\}$, $\tilde{X}_k$ 和 $\tilde{y}_k$ 的计算类似于第 1 节中的 $\tilde{x}_i$ 和 $\tilde{y}_i$. 然而式 (4) 单独使用当前数据子集并不能概括完整信息, 因此需要合并历史分析再次计算. 设 $\tilde{\beta}_{k-1}$ 是在时间点 $k-1$ 得到的在线更新估计, 则在时间点 k 对参数 β 的在线更新估计可以通过 $\tilde{\beta}_{k-1}$ 和新的数据子集来计算得到:

$$\tilde{\beta}_k = \left(V_{k-1} + \tilde{X}_k^{\top}W_k\tilde{X}_k\right)^{-1}\left(V_{k-1}\tilde{\beta}_{k-1} + \tilde{X}_k^{\top}W_k\tilde{X}_k\hat{\beta}_{k,n_k}\right) \tag{5}$$

其中, $V_k = \sum_{l=1}^{k} \tilde{X}_l^{\top}W_l\tilde{X}_l$, $k = 1,2,\cdots$, 初始值 V_0 和 $\tilde{\beta}_0$ 设置为 $V_0 = 0, \tilde{\beta}_0 = 0$.

式 (5) 的在线估计量 $\tilde{\beta}_k$ 借鉴了 Schifano 等[21]针对数据流下线性模型的在线估计方法, 目的是在部分线性模型中, 随着每次新数据块的加入, 参数估计值从 $\tilde{\beta}_{k-1}$ 到 $\tilde{\beta}_k$ 的持续更新优化. 在计算过程中, 每次只需保留上一个时刻更新计算所得统计量 $(V_{k-1}, \tilde{\beta}_{k-1})$, 并在当前更新步骤中根据最新观测数据子集计算新的统计量 $(\tilde{X}_k^{\top}W_k\tilde{X}_k, \hat{\beta}_{k,n_k})$. 式 (5) 更新估计的计算复杂度约为 $\mathrm{O}(n_k^2p+p^3)$, 空间复杂度 $\mathrm{O}(n_kp+n_kq+p^2)$. 而传统方法的计算复杂度约为 $\mathrm{O}(N_k^2p+p^3)$, 空间复杂度 $\mathrm{O}(N_kp+N_kq+p^2)$. 因此在线估计方法不仅操作简便, 而且在处理数据流的分析中计算成本更低, 能快速对参数估计持续更新.

值得注意的是, 在普通最小二乘 (OLS) 估计中, 通常假设设计矩阵是列满秩的. 同样, 在式 (4) 和式 (5) 中也需要确保可逆性假设. 然而式 (4) 可以给出另一种表达形式 $\tilde{X}_k^{\top}W_k\tilde{X}_k\hat{\beta}_{k,n_k} = \tilde{X}_k^{\top}W_k\tilde{y}_k$, 这意味着在式 (5) 中可以不直接使用 $\hat{\beta}_{k,n_k}$, 因此并不要求矩阵 $\tilde{X}_1^{\top}W_1\tilde{X}_1$, $\tilde{X}_2^{\top}W_2\tilde{X}_2$, $\cdots$, $\tilde{X}_k^{\top}W_k\tilde{X}_k$ 均可逆. 如果初始权重矩阵 $\tilde{X}_1^{\top}W_1\tilde{X}_1$ 是奇异的, 按照 Schifano 等[21]提供的方法, 可以采用类似于岭回归的策略. 具体来说, 更改初始设置为 $V_0 = \lambda I_p\ (\lambda > 0)$ 开始进行更新估计, 而不是从 $V_0 = 0$ 开始. 直到进行足够多次数的更新, 在不影响式 (5) 中正常求逆计算的情况下, 可以去除 λI_p 项.

2.2 非参数部分的在线更新估计

两步估计法通过用估计结果替换部分线性模型的参数部分, 从而将模型简化成非参数回归估计. 因此, 本文方法第 2 步的目标是对非参数部分进行在线更新估计, 即估计 $g(z|\hat{\beta})$. 在本节中, 首先将介绍使用局部常数核估计的在线更新估计量, 这是一种比较简单的方法. 然后, 为了提高估计过程的精确度和适应性, 本文将此方法扩展到基于局部线性核估计的在线估计.

2.2.1 基于局部常数核估计的在线更新估计

为简化处理, 这里首先定义几个重要的汇总统计量, 旨在有效压缩并准确概括截止至时间点 k 的累积

数据集信息. 具体定义如下所示:

$$G_k(z,y)=\sum_{l=1}^{k}\sum_{m=1}^{n_l}K_{h_l}(z_{l,m}-z)\,y_{l,m} \tag{6}$$

$$G_k(z,X)=\sum_{l=1}^{k}\sum_{m=1}^{n_l}K_{h_l}(z_{l,m}-z)\,x_{l,m} \tag{7}$$

以及

$$G_k(z)=\sum_{l=1}^{k}\sum_{m=1}^{n_l}K_{h_l}(z_{l,m}-z) \tag{8}$$

其中, 符号 $K_{h_l}(\cdot)$ 表示具有带宽 $h_l\,(>0)$ 的核函数, 即 $K_{h_l}(\cdot)=K(\cdot/h_l)/h_l^q$.

通过这些汇总统计量, 结合第 2.1 节得到的参数估计 $\tilde{\beta}_k$, 非参数部分的局部常数核估计的在线计算形式可表示为:

$$\hat{g}_k(z|\tilde{\beta}_k)=\frac{G_k(z,y)-G_k^{\top}(z,X)\tilde{\beta}_k}{G_k(z)} \tag{9}$$

式 (9) 所采用的在线估计方法与第 2.1 节式 (5) 的参数估计更新方法存在细微差别. 具体地, 式 (5) 通过保留前一时刻的估计值 $\tilde{\beta}_{k-1}$ 来更新估计 $\tilde{\beta}_k$, 但在式 (9) 中, 计算 $\hat{g}_k(z|\tilde{\beta}_k)$ 不需要保留 $\hat{g}_{k-1}(z|\tilde{\beta}_{k-1})$, 而是直接使用最新的汇总统计量及参数估计值 $\tilde{\beta}_k$ 进行计算. 此外, 更新汇总统计量的过程仅涉及将先前的汇总统计量 $\{G_{k-1}(z),G_{k-1}(z,y),G_{k-1}(z,X)\}$ 与新数据块中的统计量进行简单的加法处理. 因此, 相较于 OFFLINE 方法使用所有原始数据的重新计算, 式 (9) 的计算和存储开销均能显著降低.

2.2.2 基于局部线性核估计的在线更新估计

为了下文清晰地阐述, 记 $z_{l,m}(z)=[1,(z_{l,m}-z)^{\top}]^{\top}$. 在此基础上, 引入与前文类似的汇总统计量, 具体如下:

$$S_k(z,y)=\frac{1}{N_k}\sum_{l=1}^{k}Z_l^{\top}(z)\Omega_l(z)y_l \tag{10}$$

$$S_k(z,X)=\frac{1}{N_k}\sum_{l=1}^{k}Z_l^{\top}(z)\Omega_l(z)X_l \tag{11}$$

以及

$$S_k(z)=\frac{1}{N_k}\sum_{l=1}^{k}Z_l^{\top}(z)\Omega_l(z)Z_l(z) \tag{12}$$

其中, $Z_l(z)=[z_{l,1}(z),\cdots,z_{l,n_l}(z)]^{\top}$, $\Omega_l(z)$ 表示对角矩阵 $\mathrm{diag}\{K_{h_l}(z_{l,1}-z),\cdots,K_{h_l}(z_{l,n_l}-z)\}$. 据此, 基于局部线性核估计的在线更新可构建为:

$$\tilde{g}_k(z|\tilde{\beta}_k)=e^{\top}[S_k(z)]^{-1}\left[S_k(z,y)-S_k(z,X)\tilde{\beta}_k\right] \tag{13}$$

其中, $e=(1,0,\cdots,0)^{\top}\in\mathbb{R}^{q+1}$. 如同式 (9) 一样, 式 (13) 在线估计的计算量大约只是单个子数据集的计算, 但由于基于局部线性核估计的 OFFLINE 方法的计算复杂度非线性增长, 所以式 (13) 的计算效率提升更明显.

在核密度估计和核回归中, 带宽的选择显著影响估计的效果. 因此, 在处理部分线性模型中的 X 对 Z 的回归, y 对 Z 的回归, 以及非参数成分对 Z 的回归时, 选择合适的带宽参数尤为关键. 为了提高估计精度, 本文的方法为每个数据子集中的参数和非参数部分各自使用不同的带宽. 具体地, 在子数据集不大的情况下采用 CV (cross-validation) 方法[27], 而对规模较大的子数据集采用 DPI (direct plug-in) 方法[28].

3 数值模拟

本节是数值模拟实验部分, 目的是验证本文提出的在线估计方法的有效性, 并与 OFFLINE 方法的估计结果进行数值对比分析.

3.1 随机数生成

实验所用随机数据根据式 (1) 的部分线性模型生成. 协变量 z 为一维情况下, 设真实函数 $g(z)=2\sin(\pi z)$, 真实参数 $\beta=(2,1,2)^{\top}$, 误差项服从标准正态分布. 在此设定中, 实验分别考虑 Case 1 和 Case 2 两种情形; 而对于协变量 z 为二维的情况, 实验则采用 Case 3 的假设. 3 种情形的协变量分布详述如下.

Case 1: $x\sim N(0,\Sigma)$, 其协方差矩阵 $\Sigma=(\sigma_{ij})_{3\times3}$, $\sigma_{ii}=1$, $\sigma_{ij}=0.5\,(i\neq j)$; $z\sim N(0,1)$. 这里的 x 与 z 相互独立.

Case 2: $x^{(j)}=(z+U^{(j)})/2$, 其中 $x^{(j)}$ 表示的是协变量 x 的第 j 个分量; $z\sim U(-1,1)$; $U^{(j)}\sim U(-1,1)$, $j=1,2,3$. 这里的 x 与 z 具有相互依赖关系.

Case 3: 设定函数 $g(z_1,z_2)=2\sin(\pi z_1)+3\cos(\pi z_2)$, 真实参数 $\beta=(1,1,2)^{\top}$; 这里的随机变量 $x\sim N(0,\Sigma)$, $z\sim N(0,\mathrm{diag}(1,1))$; 误差项服从标准正态分布.

3.2 实验环境

每种情形都重复实验 s=100 次. 为了符合本文考虑的按块到达的数据流场景, 在每次重复实验中, 假设在数据流截止观测点 K, 总共收集到 K 个数据块子样本, 并且其中每个子样本的样本量 n_k 均有 $n_k=n$ $(k=1,2,\cdots,K)$. 为比较不同条件下的影响, 进行了两组模拟

实验: 第 1 组实验固定总样本量N_K不变, 通过改变子样本量的大小以观察其对在线估计效果的影响; 与之相反, 第 2 组实验则固定子样本量$n=1000$, 通过改变总样本量N_K的大小来分析更新次数与估计结果的关系.

在所有模拟计算中, 本文使用高斯核函数进行核估计. 由于涉及的部分子数据集样本量较大, 计算全程采用 DPI 法选取带宽. 所有实验均在配备 16 GB RAM 的 Intel Core i5-8265U CPU 上完成.

3.3 评估指标

为了评估在线估计方法和传统方法在参数β上的估计表现, 本文将经验均方误差 (*MSE*) 用作评估指标, 其计算方式为:

$$MSE(\breve{\beta})=s^{-1}\sum_{i=1}^{s}(\breve{\beta}_i-\beta)^2 \tag{14}$$

其中, $\breve{\beta}_i$表示的是参数β的任意估计量$\breve{\beta}$在第 i 次重复实验中的结果. 此外, 本文还使用相对效率 (RE) 进行直观比较, 即传统方法与在线方法估计结果的 *MSE* 之比.

对于非参数部分, 本文在 100 个评估点上进行计算, 所以其评估指标 *MSE* 计算为:

$$MSE(\breve{g})=\sum_{i=1}^{100}\left(\breve{g}(t_i)-g(t_i)\right)^2/100 \tag{15}$$

其中, $\breve{g}(\cdot)$是对未知函数$g(\cdot)$的任意估计. 3 种情形中函数$g(t)$的评估点分别设置如下: 在 Case 1 中, 评估点$t_i=-2+4(i-1)/99$, $i=1,2,\cdots,100$; Case 2 中, 评估点为$t_i=-1+2(i-1)/99$, $i=1,2,\cdots,100$; Case 3 中, 评估点集合为$\left\{\left(t^{(1)},t^{(2)}\right)\mid t^{(1)},t^{(2)}\in\{-1,-7/9,\cdots,1\}\right\}$, 这里的$t^{(1)}$和$t^{(2)}$是函数自变量 t 的分量.

3.4 实验结果

(1) 计算时间对比

表 1 和表 2 分别列出了在线估计方法和传统方法的平均计算时间, 其中 ONLINE 列的时间表示的是在线方法单次更新估计的计算耗时, OFFLINE 列代表传统方法更新估计的计算耗时, “LCR”和“LLR”分别代表基于局部常数核回归和局部线性核回归的在线估计方法. 表 1 和表 2 的结果表明本文的方法比起 OFFLINE 方法显著减少单次更新耗时: ONLINE 方法更新参数β估计的计算耗时约是 OFFLINE 方法的 $(n_k/N_k)^2$ 倍, 基于 LCR 在线方法更新非参数部分估计的耗时约是 OFFLINE 方法的n_k/N_k倍, 基于 LLR 在线方法更新非参数部分的耗时约是 OFFLINE 方法的 $(n_k/N_k)^2$ 倍. 由于子样本集的规模 n_k 基本不变, 耗时上的表现说明随着流数据的总样本量N_K越大, 本文的在线估计方法计算效率上的优势也就越大.

表 1 固定总样本量 N_K, 改变子样本量 n 时 ONLINE 方法和 OFFLINE 方法的计算时间对比 (s)

Case	Time	OFFLINE	ONLINE		
		N_K=20000	n=500, K=40	n=1000, K=20	n=2000, K=10
Case 1	beta.time	239.61	0.14	0.50	2.05
	LCR.time	0.28	0.01	0.02	0.03
	LLR.time	442.51	0.31	1.16	4.64
Case 2	beta.time	247.10	0.14	0.51	2.05
	LCR.time	0.32	0.01	0.02	0.03
	LLR.time	445.94	0.30	1.17	4.54

Case	Time	OFFLINE	ONLINE	
		N_K=10000	n=500, K=20	n=1000, K=10
Case 3	beta.time	657.09	1.84	6.73
	LCR.time	1.52	0.08	0.15
	LLR.time	115.89	0.35	1.25

表 2 改变总样本量 N_K, 固定子样本量 n 时 ONLINE 方法和 OFFLINE 方法的计算时间对比 (s)

Case	Time	ONLINE	OFFLINE			
		n=1000	N_K=5000	N_K=10000	N_K=15000	N_K=20000
Case 1	beta.time	0.54	13.50	60.66	135.79	239.61
	LCR.time	0.02	0.09	0.15	0.22	0.28
	LLR.time	1.25	29.50	113.35	260.02	442.51
Case 2	beta.time	0.57	14.28	59.97	139.19	247.10
	LCR.time	0.02	0.09	0.16	0.23	0.32
	LLR.time	1.13	30.87	107.89	244.21	445.94
Case 3	beta.time	6.73	164.96	657.09	—	—
	LCR.time	0.15	0.77	1.52	—	—
	LLR.time	1.25	31.49	115.89	—	—

(2) β估计的对比

表 3 和表 4 分别展示了使用 ONLINE 方法和 OFFLINE 方法估计β的经验偏差 (Bias)、经验标准误差 (SE)、经验均方误差 (*MSE*) 以及相对效率 (RE). 可以看到在估计的偏差上, 在线估计量$\tilde{\beta}_K$与 OFFLINE 估计量$\hat{\beta}$的偏差很接近. 在估计的标准误差方面, 在线估计一致表现出比 OFFLINE 方法更低的标准误差, 这可能是因为本文的在线估计方法分块进行更新估计, 因此结果具有更好的稳定性. 此外, 固定总样本量N_K, 随着单位子样本量大小的增加; 或者固定单位子样本量, 随着总更新次数 K 的增加 (如图 1), 这两种情况下$\tilde{\beta}_K$的 *MSE* 都有着稳步下降的趋势, 直到接近 OFFLINE 方法的基准, 相对效率接近 1. 表明对于参数β的估计, 本文的在线方法在均方误差上的表现接近传统方法.

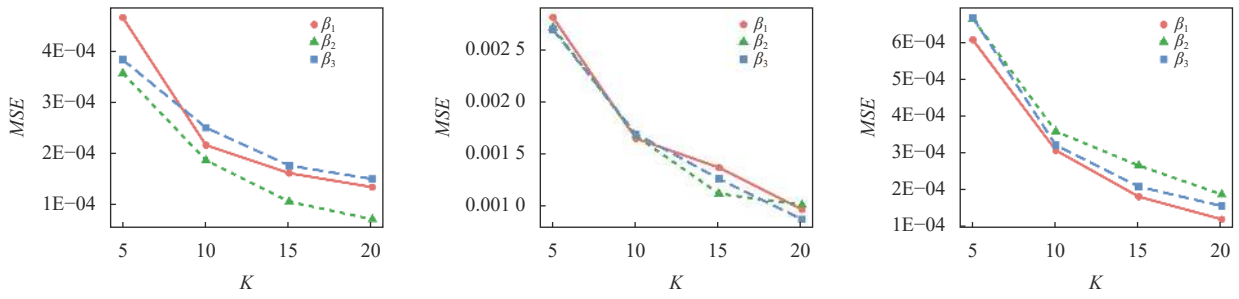
表 3 固定总样本量 N_K, 改变子样本量 n 时 ONLINE 方法和 OFFLINE 方法对参数估计的 Bias (10^{-3}), SE (10^{-3}), *MSE* (10^{-4})

Case	指标	OFFLINE			ONLINE								
		N_K=20000			n=500, K=40			n=1000, K=20			n=2000, K=10		
		$\hat{\beta}^{(1)}$	$\hat{\beta}^{(2)}$	$\hat{\beta}^{(3)}$	$\tilde{\beta}_K^{(1)}$	$\tilde{\beta}_K^{(2)}$	$\tilde{\beta}_K^{(3)}$	$\tilde{\beta}_K^{(1)}$	$\tilde{\beta}_K^{(2)}$	$\tilde{\beta}_K^{(3)}$	$\tilde{\beta}_K^{(1)}$	$\tilde{\beta}_K^{(2)}$	$\tilde{\beta}_K^{(3)}$
Case 1	Bias	−1.08	1.08	−1.42	−6.77	−1.18	−7.10	−4.92	−0.41	−5.26	−3.72	−0.10	−3.88
	SE	10.69	8.63	11.17	10.67	8.74	11.20	10.67	8.66	11.24	10.73	8.69	10.91
	MSE	1.14	0.75	1.26	1.59	0.77	1.74	1.37	0.74	1.53	1.28	0.75	1.33
	RE	—	—	—	0.72	0.97	0.72	0.84	1.01	0.82	0.89	1.00	0.94
Case 2	Bias	−17.44	−16.03	−17.34	−21.64	−16.63	−22.59	−19.99	−16.12	−19.57	−17.26	−16.18	−17.94
	SE	29.05	30.17	24.97	25.09	27.65	23.58	24.29	27.85	22.62	23.25	27.95	22.95
	MSE	11.40	11.58	9.18	10.91	10.33	10.60	9.84	10.27	8.90	8.33	10.35	8.43
	RE	—	—	—	1.04	1.12	0.87	1.16	1.13	1.03	1.37	1.12	1.09

Case	指标	OFFLINE			ONLINE					
		N_K=10000			n=500, K=20			n=1000, K=10		
		$\hat{\beta}^{(1)}$	$\hat{\beta}^{(2)}$	$\hat{\beta}^{(3)}$	$\tilde{\beta}_K^{(1)}$	$\tilde{\beta}_K^{(2)}$	$\tilde{\beta}_K^{(3)}$	$\tilde{\beta}_K^{(1)}$	$\tilde{\beta}_K^{(2)}$	$\tilde{\beta}_K^{(3)}$
Case 3	Bias	−1.60	−0.37	0.38	−1.61	−1.36	2.19	−2.19	−0.34	1.42
	SE	16.78	18.11	16.82	19.39	19.58	18.93	17.57	19.09	18.02
	MSE	2.81	3.25	2.80	3.75	3.81	3.60	3.10	3.61	3.24

表 4 改变总样本量 N_K, 固定子样本量 n=1000 时 ONLINE 方法对参数估计的 Bias (10^{-3}), SE (10^{-3}), *MSE* (10^{-4})

Case	指标	n=1000, K=5			n=1000, K=10			n=1000, K=15			n=1000, K=20		
		$\tilde{\beta}_K^{(1)}$	$\tilde{\beta}_K^{(2)}$	$\tilde{\beta}_K^{(3)}$	$\tilde{\beta}_K^{(1)}$	$\tilde{\beta}_K^{(2)}$	$\tilde{\beta}_K^{(3)}$	$\tilde{\beta}_K^{(1)}$	$\tilde{\beta}_K^{(2)}$	$\tilde{\beta}_K^{(3)}$	$\tilde{\beta}_K^{(1)}$	$\tilde{\beta}_K^{(2)}$	$\tilde{\beta}_K^{(3)}$
Case 1	Bias	−6.39	−0.14	−5.61	−5.94	−0.73	−4.13	−5.12	−0.80	−4.99	−4.92	−0.41	−5.26
	SE	20.72	19.01	18.87	13.60	13.80	15.42	11.80	10.45	12.47	10.67	8.66	11.24
	MSE	4.66	3.58	3.84	2.18	1.89	2.52	1.64	1.09	1.79	1.37	0.74	1.53
	RE	0.89	0.95	0.97	0.83	0.96	0.94	0.87	1.00	0.88	0.84	1.01	0.82
Case 2	Bias	−19.60	−12.40	−18.70	−18.62	−18.15	−21.53	−19.75	−16.43	−20.03	−19.99	−16.12	−19.57
	SE	49.49	50.87	48.66	36.39	36.95	35.39	31.65	29.54	29.69	24.29	27.85	22.62
	MSE	28.09	27.16	26.94	16.58	16.81	17.03	13.82	11.34	12.74	9.84	10.27	8.90
	RE	0.91	1.00	0.95	0.89	1.20	0.95	1.72	1.31	1.12	1.16	1.13	1.03
Case 3	Bias	−1.86	0.66	1.24	−2.19	−0.34	1.42	−2.64	−0.64	1.20	−1.93	−0.20	−0.04
	SE	24.71	25.91	25.95	17.57	19.09	18.02	13.39	16.48	14.59	11.02	13.90	12.70
	MSE	6.08	6.65	6.68	3.10	3.61	3.24	1.85	2.69	2.12	1.24	1.91	1.60



(a) Case 1 (b) Case 2 (c) Case 3

图 1 子样本量 n=1000 时, ONLINE 方法参数估计的 *MSE* 随更新次数 K 变化趋势

(3) 函数$g(\cdot)$估计的对比

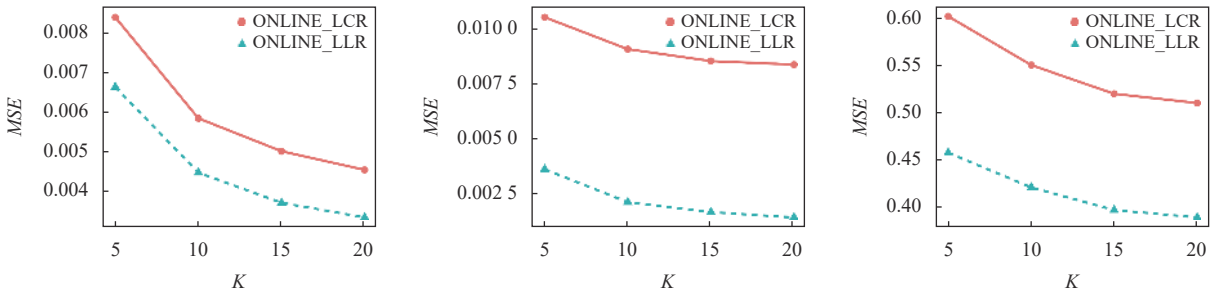
表 5 和表 6 给出了两种方法对非参数部分估计的实验结果. 表中列出 100 次重复实验的偏差平方、方差以及经验均方误差的平均值. 尽管在线估计量的偏差平方大多略高于 OFFLINE 方法, 但其值仍然较低. 此外, 估计结果的方差均低于 OFFLINE 方法. 结果表明, 更大的单位子样本量虽然有助于减少偏差, 但是会导致方差的增大. 而在相同的单位子样本量下, 随着更新次数的增加, 偏差平方和方差都有所降低. 总体而言, 本文的在线方法通过多次更新, 能够逐渐降低估计非参数部分的 *MSE*, 如图 2 所示. 值得注意的是, 在实验的所有情形中, 基于 LLR 的在线估计方法相较于基于 LCR 的在线估计方法具有更低的 *MSE*, 其在经验均方误差上的表现更接近于 OFFLINE 方法, 表明在偏差和方差之间的平衡上, 该方法具有更好的控制性能.

表 5 固定总样本量 N_K, 改变子样本量 n 时 ONLINE 和 OFFLINE 方法对函数 $g(\cdot)$ 估计的 Bias^2 (10^{-3}), Var (10^{-3}), *MSE* (10^{-3})

Case	指标	OFFLINE		ONLINE					
		N_K=20000		n=500, K=40		n=1000, K=20		n=2000, K=10	
		LCR	LLR	LCR	LLR	LCR	LLR	LCR	LLR
Case 1	Bias^2	0.71	0.45	9.58	6.37	6.01	3.98	3.67	2.41
	Var	3.22	3.42	1.84	1.71	1.98	1.86	2.19	2.08
	MSE	3.93	3.87	11.42	8.08	8.00	5.84	5.86	4.49
Case 2	Bias^2	1.87	0.09	11.02	1.39	7.78	0.77	5.43	0.42
	Var	1.01	2.42	0.58	0.64	0.62	0.68	0.67	0.73
	MSE	2.88	2.51	11.60	2.02	8.40	1.45	6.10	1.15
Case	指标	OFFLINE		ONLINE					
		N_K=10000		n=500, K=20		n=1000, K=10			
Case 3	Bias^2	466.67	371.11	545.22	415.17	530.91	407.60		
	Var	124.84	133.71	13.50	9.20	19.23	13.85		
	MSE	591.51	504.82	558.72	424.37	550.14	421.46		

表 6 改变总样本量 N_K, 固定子样本量 n=1000 时 ONLINE 方法对函数 $g(\cdot)$ 估计的 Bias^2 (10^{-3}), Var (10^{-3}), *MSE* (10^{-3})

Case	指标	n=1000, K=5		n=1000, K=10		n=1000, K=15		n=1000, K=20	
		LCR	LLR	LCR	LLR	LCR	LLR	LCR	LLR
Case 1	Bias^2	3.93	2.4	3.67	2.41	3.58	2.34	3.47	2.32
	Var	4.48	4.25	2.19	2.08	1.46	1.38	1.08	1.03
	MSE	8.41	6.65	5.86	4.49	5.03	3.73	4.56	3.36
Case 2	Bias^2	8.06	0.93	7.84	0.79	7.71	0.76	7.78	0.77
	Var	2.49	2.71	1.26	1.35	0.86	0.93	0.62	0.68
	MSE	10.55	3.64	9.1	2.14	8.56	1.69	8.4	1.45
Case 3	Bias^2	556.43	425.69	530.91	407.6	507.97	389.09	501.13	383.36
	Var	45.06	32.22	19.23	13.85	12	8.58	9.1	6.51
	MSE	601.49	457.91	550.14	421.46	519.97	397.66	510.22	389.87

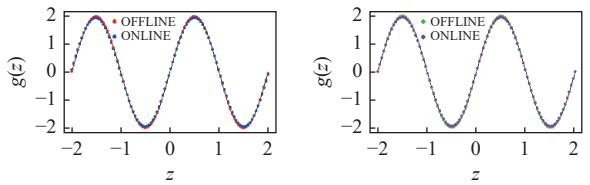


(a) Case 1 (b) Case 2 (c) Case 3

图 2 子样本量 n=1000 时, ONLINE 方法对非参数部分估计的 *MSE* 随更新次数 K 变化趋势

图 3 进一步展示 Case 1 和 Case 2 中非参数部分估计的图形. 其中, 黑色实线表示真正的函数形式, ONLINE 方法和 OFFLINE 方法的估计以不同颜色的圆点代表. 基于 LCR 的在线估计方法主要劣势在于拟合边界处表现出更多的边缘效应, 但其在内部点上的估计基本上接近 OFFLINE 方法的估计. 相比之下, 基于 LLR 的在线估计在边界和内部点上均表现优异, 几乎与传统方法的估计完全重叠. 图 4 展示了在 Case 3 中, 当 z 为二维时的估计图形, 其中, 蓝色“+”表示 OFFLINE 方法的估计, 红色圆点表示 ONLINE 方法的估计. 基于 LCR 和 LLR 的在线方法仍然表现出与 OFFLINE 方法几乎一致的估计结果.

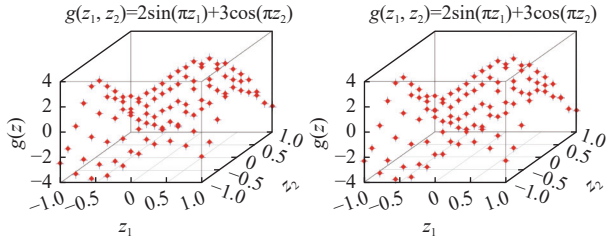
综上所述, 通过考虑协变量之间的不同关系以及 z 为一维和二维情况下的模拟实验, 本文的方法不仅表现出估计准确性和稳定性, 同时相较于传统方法具有显著的计算效率优势.



(a) Case 1: LCR (b) Case 1: LLR

(c) Case 2: LCR (d) Case 2: LLR

图 3 当总样本量 N_K=20000, 子样本量 n=1000 时, ONLINE 和 OFFLINE 方法对未知函数 $g(z)=2\sin(\pi z)$的逐点估计



(a) Case 3: LCR (b) Case 3: LLR

图 4 当总样本量 N_K=10000, 子样本量 n=1000 时, ONLINE 方法和 OFFLINE 方法对 Case 3 中未知函数的逐点估计

4 实际数据分析

本节应用本文的在线估计方法对我国劳动年龄人口的生活质量进行实际数据分析, 研究相关因素对生活质量的影响, 以及提高生活品质的可行途径.

4.1 数据预处理

基于 CGSS 数据集, 本文采用 2017 年、2018 年、2021 年最近 3 次的调查数据以及时反映当下的社会情况, 聚焦于处于社会中流砥柱的劳动年龄人口, 根据国际标准将 18–65 岁年龄段的样本筛选出来. 进一步地, 将其中仍在上学阶段和已经离/退休的人群排除, 数据清洗后总共获得 17495 个有效样本.

(1) 因变量

由于本文关注的对象是现今劳动年龄群体的生活质量, 旨在真实反映人们的获得感、幸福感、安全感. 因此, 这里将主观生活质量综合指标设定为因变量. 该指标基于个体对身体健康、心理健康、幸福程度、社会地位和经济水平 5 个方面的自我评价构建而成, 分别对应于问卷中的 A15、A17、A36、A43a 和 A43e 项. 在此基础上, 将这些指标视为同等重要的变量. 鉴于 A43a 的原始数据采用十分制, 而其他变量采用的是五分制, 所以在计算中 A43a 项的权重调整为其他变量的一半. 特别地, 由于原始数据中 A43e 项将经济地位阶层由高到低标记为 1–5, 所以本文将其重新赋值, 以 1–5 表示从低到高的经济地位. 通过采用 min-max 标准归一化处理 5 项变量加权求和的结果, 并将其转换为百分制以直观展示生活质量的综合得分, 使用记号“score”来表示该因变量.

相关变量的描述统计见表 7, 频数分布见图 5, 图示生活质量综合得分 (score) 的频数分布图总体呈现钟形分布, 生活质量极高与极低的频率均较低, 样本中大多数劳动年龄群体的生活质量集中在中等水平, 反映出样本总体的生活质量分布较为均衡.

表 7 生活质量综合指标相关变量的描述统计

变量名	指标	均值	标准差	解释
A15	自评身体健康程度	3.65	1.06	1–5定序变量
A17	自评心理健康程度	3.84	0.99	1–5定序变量
A36	自评幸福程度	3.84	0.84	1–5定序变量
A43a	自评社会地位阶层	4.16	1.69	1–10定序变量
A43e	自评经济地位阶层	2.25	0.86	1–5定序变量
score	生活质量综合得分	54.46	14.85	0–100连续变量

(2) 解释变量

选择合适的解释变量, 对于分析影响生活质量的相关可能尤为重要, 所幸 CGSS 提供了全面广泛的社会信息. 本文通过对问卷回答的分析, 从人口属性、家庭环境、生活方式、工作环境、社会保障、经济收入 6 大因素来选择解释变量. 具体变量描述统计见表 8, 并在图 6 展示变量之间的相关系数矩阵图.

(3) 基本分析

首先, 本文通过图 7 的散点图辅助分析上述若干解释变量中的连续变量与因变量之间的关系: 其中, 变量 age、height、log_income 与 score 之间的散点图呈现出大体上的线性关系, 但在周工作时间与 score 的散点图上, 拟合曲线显示出不规则的波动, 表明工作时间与生活质量之间存在可能的非线性关系. 结合图 6 显示各变量之间的线性相关性较弱, 因此, 本文决定采用部分线性模型进行建模, 并将周工作时间作为模型中的非线性因素, 其余变量则纳入模型的线性部分进行分析.

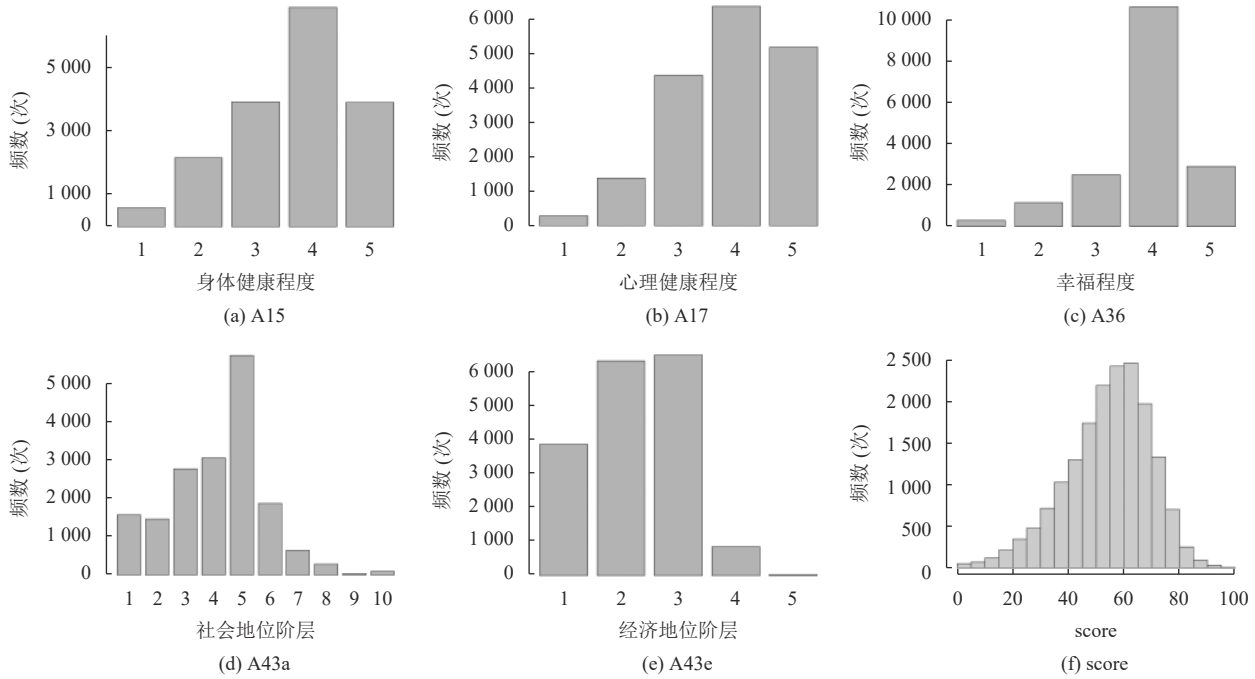


图 5　生活质量综合指标相关变量的频数分布图

表 8　解释变量的描述统计

变量名	指标	均值	标准差	解释
age	年龄	44.19	11.99	18–65整数
edu	学历	5.44	3.29	1–13定序变量
height	身高	164.69	8.00	单位为厘米
bmi	肥胖程度	2.38	0.78	1–5定序变量
hukou	户口类别	0.30	0.46	非农=1, 农业=0
marital_status	婚姻状态	0.81	0.39	有配偶=1, 无配偶=0
sport	运动频率	2.50	1.54	1–5定序变量
socialize	社交频率	4.04	1.78	1–7定序变量
medicare	是否参与医保	0.93	0.26	参加=1, 未参加=0
log_income	人均收入水平	9.59	1.83	人均收入的对数
work_time	每周工作时间	35.44	27.93	单位为小时

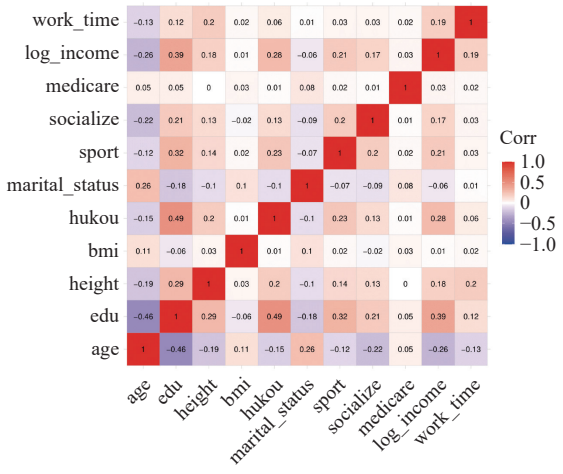


图 6　变量之间相关系数矩阵图

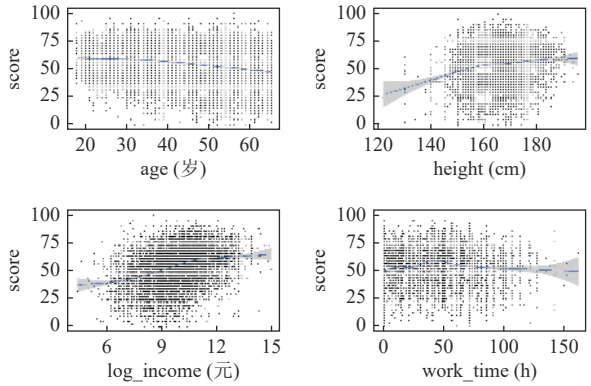


图 7　连续变量与因变量之间关系的散点图

其次, 考虑到部分变量是定序变量, 本文进一步对包括定序变量在内的相关变量与因变量之间的线性关系进行了逐一检验 (见表 9). 结果表明, 除了 bmi 变量的线性影响不显著外, 其余变量的线性回归估计均通过了检验. 据此, 在后续的回归分析中, 本文将 bmi 变量通过哑变量处理后纳入模型, 而将其他的定序变量按原有的序列性质直接纳入模型的线性部分, 避免不必要的数据转换.

4.2　模型构建

基于第 4.1 节的数据预处理, 本节将构建部分线性模型, 以准确反映各因素对生活质量的影响, 并应用先前提出的大数据在线估计方法进行分析. 模型以主观

生活质量综合得分 (score) 作为响应变量, 选取周工作时间作为非线性部分的协变量. 对于模型的线性部分, 协变量 X_1 表示年龄; X_2 表示为学历水平; X_3 表示为身高; 并将 bmi 变量转换为哑变量, 以身体质量正常组为对照组, X_4、X_5、X_6 和 X_7 分别代表消瘦、过重、肥胖和过度肥胖组. 其他线性协变量包括: X_8 户口类别, X_9 婚姻状态, X_{10} 运动频率, X_{11} 社交频率, X_{12} 参与医保状态, 以及 X_{13} 人均年家庭收入的对数.

表 9 各变量逐一与因变量的线性关系检验

变量名	Estimate	Std. Error	t value	Pr(>|t|)
age	−0.2944	0.0091	−32.37	<2E−16
height	0.2909	0.0139	21.00	<2E−16
log_income	2.4783	0.0584	42.44	<2E−16
edu	1.3728	0.0326	42.17	<2E−16
sport	2.1882	0.0713	30.71	<2E−16
socialize	1.7903	0.0617	29.00	<2E−16
bmi	−0.1547	0.1443	−1.07	0.284

视 2017–2021 年间 3 次问卷调查数据为流数据, 假设数据分批次到达, 单位子数据集样本量$n = 500$. 则总样本量$N_K = 17495$的样本需要使用本文的方法在线更新估计$K = 35$次. 鉴于子样本量$n = 500$规模相对适中, 带宽选择采用 CV 方法.

4.3 估计结果及解释

(1) 参数部分估计结果

本文的在线估计方法进行更新计算 35 次后, 得到参数估计结果如表 10 所示.

表 10 模型参数的在线估计结果

变量	系数	变量	系数
X_1 (age)	−0.21	X_8 (hukou)	0.52
X_2 (edu)	0.54	X_9 (marital_status)	5.24
X_3 (height)	0.04	X_{10} (sport)	1.28
X_4 (underweight)	−2.47	X_{11} (socialize)	1.12
X_5 (overweight)	0.06	X_{12} (medicare)	0.64
X_6 (obesity)	−1.13	X_{13} (log_income)	1.04
X_7 (severe obesity)	−1.22		

表 10 的估计结果表明, 健康因素对生活质量的影响尤为显著. 与健康直接相关的变量, 如参与医保、适中的 bmi 指数、规律运动和社交活动均与生活质量显示出正相关性, 这些因素共同凸显了保持良好健康状况对于提高生活质量的重要性. 此外, 随着年龄的增长, 生活质量可能面临下降的风险, 社会支持和职场环境需更加关注中老年人群的特定需求, 如健康维护、职业培训及心理健康服务.

教育程度和婚姻状况对劳动年龄人口的生活质量呈正相关. 这在一定程度上说明, 提升教育水平和支持家庭稳定性不仅有助于促进个人职业发展和生活满意度的提高, 也有助于构建更和谐文明的社会环境.

而在经济因素上, 结果表现出与劳动年龄人口的生活质量密切相关. 家庭年收入的增加和城市户口直接关联到生活质量的提升. 推动经济发展是改善生活质量的有效途径, 但同时也需要重视资源公平分配, 确保所有群体尤其是低收入和边缘群体能够获得提升生活质量的机会.

(2) 非参数部分估计结果

在非参数部分的估计结果如图 8 所示. 曲线趋势表明, 当周工作时间超过 60 h, 生活质量开始显著下降, 而超过 100 h 后, 生活质量的下降幅度加剧, 呈现出明显的断崖式下跌. 这种趋势暗示长时间工作可能引发健康或家庭等方面的问题, 从而对生活质量产生负面影响. 另一方面, 当周工作时间少于 30 h, 生活质量亦呈现下滑趋势, 这可能与经济收入减少、职业发展受限等因素有关.

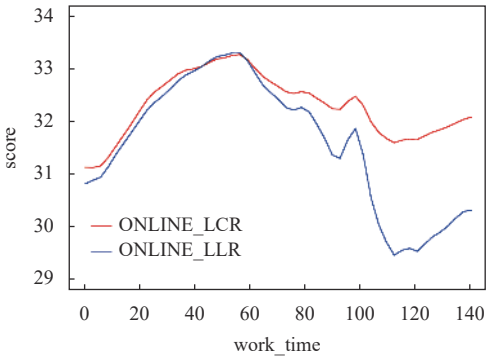


图 8 周工作时间对生活质量 (score) 的影响曲线

而生活质量最高的群体是周工作时间在 30–60 h 之间的全职工作者. 一份稳定的全职工作不仅提供了必要的经济支持, 也带来了个人生活满意度的提升. 基于这一发现, 建议社会政策制定应关注失业和不完全就业的问题, 推广灵活且科学的工作时间配置, 以确保劳动者能够在保持生产力的同时, 享有足够的休息和个人时间, 创造一个健康、高效的工作环境.

5 总结

本文提出了一种在线估计方法, 专为数据流场景下的部分线性模型而设计. 这一方法的核心优势在于其快速的计算能力和对数据存储需求的大幅降低, 有效地解决了部分线性模型大数据分析的重点难题. 经过数值模拟验证, 本文的方法展现了高效的计算能力,

并在均方误差方面的表现与传统方法相当, 是一种快速且有效的数据流计算方法, 为大数据时代下的部分线性模型分析提供了新的视角和工具.

参考文献

1 Engle RF, Granger CWJ, Rice J, *et al*. Semiparametric estimates of the relation between weather and electricity sales. Journal of the American Statistical Association, 1986, 81(394): 310–320. [doi: 10.1080/01621459.1986.10478274]

2 Robinson PM. Root-*N*-consistent semiparametric regression. Econometrica, 1988, 56(4): 931–954. [doi: 10.2307/1912705]

3 Speckman P. Kernel smoothing in partial linear models. Journal of the Royal Statistical Society: Series B (Methodological), 1988, 50(3): 413–436. [doi: 10.1111/j.2517-6161.1988.tb01738.x]

4 Auerbach E. Identification and estimation of a partially linear regression model using network data. Econometrica, 2022, 90(1): 347–365. [doi: 10.3982/ECTA19794]

5 Heckman NE. Spline smoothing in a partly linear model. Journal of the Royal Statistical Society: Series B (Methodological), 1986, 48(2): 244–248. [doi: 10.1111/j.2517-6161.1986.tb01407.x]

6 Chen H. Convergence rates for parametric components in a partly linear model. The Annals of Statistics, 1988, 16(1): 136–146. [doi: 10.1214/aos/1176350695]

7 Wang QH, Linton O, Härdle W. Semiparametric regression analysis with missing response at random. Journal of the American Statistical Association, 2004, 99(466): 334–345. [doi: 10.1198/016214504000000449]

8 Liang H, Wang SJ, Robins JM, *et al*. Estimation in partially linear models with missing covariates. Journal of the American Statistical Association, 2004, 99(466): 357–367. [doi: 10.1198/016214504000000421]

9 Liang H, Wang S, Carroll RJ. Partially linear models with missing response variables and error-prone covariates. Biometrika, 2007, 94(1): 185–198. [doi: 10.1093/biomet/asm010]

10 Ma P, Mahoney MW, Yu B. A statistical perspective on algorithmic leveraging. The Journal of Machine Learning Research, 2015, 16(1): 861–911.

11 Wang HY, Zhu R, Ma P. Optimal subsampling for large sample logistic regression. Journal of the American Statistical Association, 2018, 113(522): 829–844. [doi: 10.1080/01621459.2017.1292914]

12 Wang HY, Yang M, Stufken J. Information-based optimal subdata selection for big data linear regression. Journal of the American Statistical Association, 2019, 114(525): 393–405. [doi: 10.1080/01621459.2017.1408468]

13 Wang HY, Ma YY. Optimal subsampling for quantile regression in big data. Biometrika, 2021, 108(1): 99–112. [doi: 10.1093/biomet/asaa043]

14 Ma P, Chen YK, Zhang XL, *et al*. Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. The Journal of Machine Learning Research, 2022, 23(1): 177.

15 Lin N, Xi RB. Aggregated estimating equation estimation. Statistics and Its Interface, 2011, 4(1): 73–83. [doi: 10.4310/SII.2011.v4.n1.a8]

16 Chen XY, Xie MG. A split-and-conquer approach for analysis of extraordinarily large data. Statistica Sinica, 2014, 24(4): 1655–1684. [doi: 10.5705/ss.2013.088]

17 吴梅红, 郭佳盛, 鞠颖, 等. 基于分层筛选和动态更新的并行选择集成算法. 计算机科学, 2017, 44(1): 48–52. [doi: 10.11896/j.issn.1002-137X.2017.01.009]

18 臧艳辉, 席运江, 赵雪章. 基于 MapReduce 的分治 k 均值聚类方法. 计算机工程与设计, 2020, 41(5): 1345–1351. [doi: 10.16208/j.issn1000-7024.2020.05.022]

19 Zhao TQ, Cheng G, Liu H. A partially linear framework for massive heterogeneous data. The Annals of Statistics, 2016, 44(4): 1400–1437. [doi: 10.1214/15-AOS1410]

20 Lian H, Zhao KF, Lv SG. Projected spline estimation of the nonparametric function in high-dimensional partially linear models for massive data. The Annals of Statistics, 2019, 47(5): 2922–2949. [doi: 10.1214/18-AOS1769]

21 Schifano ED, Wu J, Wang C, *et al*. Online updating of statistical inference in the big data setting. Technometrics, 2016, 58(3): 393–403. [doi: 10.1080/00401706.2016.1142900]

22 Lee J, Wang HY, Schifano ED. Online updating method to correct for measurement error in big data streams. Computational Statistics & Data Analysis, 2020, 149: 106976. [doi: 10.1016/j.csda.2020.106976]

23 Luo L, Song PXK. Renewable estimation and incremental inference in generalized linear models with streaming data sets. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2020, 82(1): 69–97. [doi: 10.1111/rssb.12352]

24 Luo L, Song PXK. Multivariate online regression analysis with heterogeneous streaming data. Canadian Journal of Statistics, 2023, 51(1): 111–133. [doi: 10.1002/cjs.11667]

25 Kong EF, Xia YC. On the efficiency of online approach to nonparametric smoothing of big data. Statistica Sinica, 2019, 29(1): 185–201. [doi: 10.5705/ss.202015.0365]

26 Li Q. On the root-*N*-consistent semiparametric estimation of partially linear models. Economics Letters, 1996, 51(3): 277–285. [doi: 10.1016/0165-1765(96)00821-X]

27 Härdle W, Liang H, Gao JT. Partially Linear Models. Berlin: Springer, 2000.

28 Ruppert D, Sheather SJ, Wand MP. An effective bandwidth selector for local least squares regression. Journal of the American Statistical Association, 1995, 90(432): 1257–1270. [doi: 10.1080/01621459.1995.10476630]

(校对责编: 孙君艳)