

基于纵向数据半参数混合效应模型的网络营销影响因素研究

汪轲頔, 毛 围

贵州大学数学与统计学院, 贵州 贵阳

收稿日期: 2024年9月18日; 录用日期: 2024年10月16日; 发布日期: 2024年11月18日

摘 要

本文建立了一个纵向数据的半参数混合效应模型, 采用B样条方法建立纵向指标与时间的关系, 为模型中的所有未知参数和随机效应指定先验分布, 在共轭先验条件下, 对该模型进行贝叶斯分析, 通过Gibbs抽样方法获取未知参数、随机效应和非参数函数的贝叶斯估计。最后将该模型应用于饮食类非遗产品网络营销的影响因素进行分析, 实证结果显示本文建立的指标对淘宝店铺的网络销售量产生了不同程度的影响, 证明了模型的有效性。

关键词

纵向数据, B样条, 贝叶斯估计, 网络营销

Research on the Influencing Factors of Online Marketing Based on Longitudinal Data Semi Parametric Mixed Effects Model

Yundi Wang, Wei Mao

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Sep. 18th, 2024; accepted: Oct. 16th, 2024; published: Nov. 18th, 2024

Abstract

This article establishes a semi parametric mixed effects model for longitudinal data, using the B-spline method to establish the relationship between longitudinal indicators and time. Prior distributions are specified for all unknown parameters and random effects in the model, and Bayesian analysis of the model is considered under conjugate priors. Bayesian estimates of unknown

parameters, random effects, and non parametric functions are extracted based on Gibbs sampling, and the model is applied to the study of influencing factors in online marketing. The empirical results show that the indicators established in this article have varying degrees of impact on the online sales volume of Taobao stores, proving the effectiveness of the model.

Keywords

Longitudinal Data, B-Spline, Bayesian Estimation, Online Marketing

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

网络营销是指利用互联网及其相关技术,通过各种在线渠道和工具进行产品或服务的推广、销售和品牌建设的过程。它包括但不限于搜索引擎优化(SEO)、搜索引擎营销(SEM)、社交媒体营销、电子邮件营销、内容营销和在线广告等策略。网络营销的目标是吸引潜在客户、提高品牌知名度、促进销售转化,并与消费者建立长期关系。通过数据分析和用户反馈,企业能够实时调整营销策略,以适应市场变化和消费者需求。在大数据网络时代,网络营销面临着前所未有的变革与机遇。目前,主要的网络营销方式包括搜索引擎营销、微信营销、数据库营销和病毒式营销等[1]-[4]。

纵向数据是指在不同时间点上对同一组个体或观察单位进行重复测量的数据。这种数据类型常用于研究时间效应、个体变化和动态过程,能够揭示个体随时间的变化趋势。在大数据背景下,纵向数据因其特点在网络营销过程中占据了重要地位。例如,在收集某搜索引擎平台上黄金的销售数据时,这些数据不仅受到黄金公司自身在产品质量、样式和促销等方面的影响,还受到国家对黄金价格的调控政策的影响。如果仅依赖截面数据,虽然可以分析各黄金公司受到多种变量的不同影响,但无法揭示国家调控政策对黄金销售的具体影响。而如果仅使用时间序列数据,则只能分析黄金价格和销售情况随时间的变化,而无法有效体现不同公司在黄金销售中受到的具体变量影响[5]。

对于纵向数据通常采用混合效应模型建模,混合效应模型中包含固定效应和随机效应。而线性混合效应模型在一些情况中拟合效果较差。因此,许多统计学家开始采用半参数混合效应模型对纵向数据进行建模。半参数混合效应模型兼具参数模型和非参数模型的优点,在有效利用数据信息的同时考虑了信息不完整的变量,增强了模型的灵活性。Angelo 等[6]通过引入 B 样条对光滑样条模型进行了优化,提出了一种新的方法以拟合非线性混合效应模型。阙焯等[7]采用广义最小二乘法对未知参数、随机效应及方差成分进行了估计,并证明了所提出估计量的渐进性质。为提高估计的效率,Lindley 等[8]将贝叶斯方法应用于线性混合模型中,而齐培艳等[9]则研究了具有变点的半参数非线性混合效应模型的多重估计方法。

半参数混合效应模型中的参数估计通常有两种方法,一种是频率学派的极大似然方法,一种是贝叶斯学派的 MCMC (Markov Chain Monte Carlo)方法。与频率学派方法相比,贝叶斯方法建模时最重要的优点就是不需要渐进逼近,由于所有参数均为随机变量,因此可以采用一些基于抽样的方法(例如蒙特卡洛方法)来构建复杂模型。同时,贝叶斯方法可以有效减轻计算负担,使得模型求解更容易[10]。

本文使用半参数混合效应模型对纵向数据进行建模,模型中纵向指标与时间的关系采用 B 样条逼近,并对模型中的所有未知参数、随机效应给出先验分布,基于贝叶斯 MCMC 方法对模型中的未知参数进行估计。最后将所提出的方法运用于网络营销影响因素的研究中,从而验证模型的有效性。

2. 模型

2.1. 半参数混合效应模型

假设有 n 个个体, 每个个体之间相互独立, 对每个个体分别进行 m_i 次观测, 本文采用以下半参数混合效应模型:

$$Y_{ij} = X_{ij}^T \beta + Z_{ij}^T b_i + f(t_{ij}) + \varepsilon_{ij} \quad (1)$$

其中, $i=1, \dots, n$; $j=1, \dots, m_i$; $\sum_{i=1}^n m_i = M$; $Y_i(t_{ij})$ 为第 i 个个体在第 j 个观测时间 t_{ij} 观察到的响应变量;

$X_i^T(t_{ij})$ 和 $Z_i^T(t_{ij})$ 分别是固定效应 β 和随机效应 b_i 的设计向量, 固定效应 $\beta = (\beta_1, \dots, \beta_p)^T$, 随机效应 b_i 服从多元正态分布 $N_q(0, D)$ 且相互独立; $f(t_{ij})$ 是一个未知的光滑函数; $\varepsilon_i(t_{ij})$ 是第 i 个个体在第 j 个观察点 t_{ij} 的测量误差, 假定 $\varepsilon_i(t_{ij})$ 服从正态分布 $N(0, \sigma^2)$ 且与 b_i 相互独立。

2.2. B 样条

参考 Ruppert 等[11], 我们在贝叶斯框架下使用 B 样条来逼近模型(1)中的未知光滑函数。具体而言, 我们考虑以下形式的光滑函数:

$$f(t) = \lambda_0^{(1)} + \lambda_1^{(1)}t + \dots + \lambda_s^{(1)}t^s + \sum_{l=1}^L \lambda_l^{(2)}(t - \tau_l)_+^s \quad (2)$$

其中, L 为光滑函数的节点数, s 为样条的自由度, 将光滑函数划分为 $L+1$ 个区间, 我们考虑按照以下规则来选取节点和样条: 假设 τ_l 为第 l 个节点, 则选取样本的第 $l+1/L+2$ 分位数, 并满足

$$\begin{aligned} \tau_0 = t_{\min} < \tau_1 < \dots < \tau_L = t_{\max}, \quad (t - \tau_l)_+^s = (\max\{0, t - \tau_l\})^s, \quad \text{令} \\ \lambda = (\lambda^{(1)T}, \lambda^{(2)T})^T = (\lambda_0^{(1)}, \lambda_1^{(1)}, \dots, \lambda_s^{(1)}, \lambda_1^{(2)}, \dots, \lambda_L^{(2)})^T, \quad \text{截断的幂基函数} \\ \varphi(t) = (\varphi_1(t)^T, \varphi_2(t)^T)^T = (1, t, \dots, t^s, (t - \tau_1)_+^s, \dots, (t - \tau_L)_+^s)^T, \quad \text{则式(2)可写为:} \end{aligned}$$

$$f(t) = \varphi_1(t)^T \lambda^{(1)} + \varphi_2(t)^T \lambda^{(2)} = \varphi(t)^T \lambda$$

记 $Y_i = (Y_{i1}, \dots, Y_{im_i})^T$, $Y = (Y_1^T, \dots, Y_n^T)^T$, 以同样方式定义 X, Z, x, ε , 设 $Z = \text{diag}(Z_1, \dots, Z_n)$ 为对角元素为 $Z_1, \dots, Z_n$ 的对角矩阵, 则式(1)可以表示为:

$$Y = X_* \beta_* + Zb + Wa + \varepsilon \quad (3)$$

其中, $X_* = (X, \varphi_1^T)$, $W = \varphi_2^T(t_{ij})_{M \times L}$, $a = \lambda^{(2)}$, $\beta_* = (\beta^T, \lambda^{(1)T})^T$, $\lambda = (\lambda^{(1)T}, \lambda^{(1)T})^T \sim N_{s+1+L}(0, \xi)$, $\xi = \text{diag}(\xi_1, \xi_2)$, $\xi_1 = \text{diag}(\sigma_{\varepsilon_1}^2, \dots, \sigma_{\varepsilon_1}^2)$, $\xi_2 = \text{diag}(\sigma_{\varepsilon_2}^2, \dots, \sigma_{\varepsilon_2}^2)$ 。

3. 贝叶斯分析

令 $\theta = \{\beta_*, D, \sigma^2, \xi\}$ 为未知参数的集合, 为了对所有未知参数进行贝叶斯估计, 需要从所有未知参数的联合似然函数中获得样本观测, 由(3)式可得似然表达式如下:

$$L(Y, b, a | \theta) = (2\pi\sigma^2)^{-n/2} \times \exp\left\{-\left(2\sigma^2\right)^{-2} \times (Y - X_* \beta_* - Zb - Wa)^T \times (Y - X_* \beta_* - Zb - Wa)\right\} \quad (4)$$

3.1. 先验的设置

联合模型的估计可以通过两种方式进行, 一种是频率学派的极大似然方法, 另一种是贝叶斯学派的 MCMC 方法。与频率学派方法相比, 贝叶斯方法建模时最重要的优点就是不需要渐进逼近, 且因为所有的参数都是随机变量, 因此可以使用基于抽样的方法(如蒙特卡洛方法)来构建复杂模型。本文使用贝叶斯

推理方法, 基于指定先验分布的联合似然, 通过马尔可夫链蒙特卡罗过程在完全贝叶斯框架下同时估计联合模型的参数。在贝叶斯框架下, 为模型中的所有未知参数和随机效应指定先验分布如下:

$$\beta_* \sim N_{p+2}(\phi_1, \phi_2), b_i \sim N_q(0, D), a \sim N_{r-2}(0, \tau I), \\ D \sim IW_q(o_o, R_o), \tau \sim IW(o_1, R_1), \sigma^2 \sim 1/\sigma^2$$

其中, IW 表示逆 Wishart 分布, σ^2 采用扩散先验分布, $\phi_1, \phi_2, o_o, R_o, o_1, R_1$ 是超参数。

3.2. 条件后验的推导

(1) 为固定效应 β_* 指定先验分布为 $\beta_* \sim N_{p+2}(\phi_1, \phi_2)$, 根据似然表达式(4), 通过多元配方法[12], 推导出 β_* 的条件后验分布:

$$\pi(\beta_* | Y, b, a, \sigma^2) \propto \exp \left\{ -\frac{1}{2} (\beta_* - (\phi_1 - \beta_*))^T \left(\frac{1}{\sigma^2} X_*^T X_* + \phi_2^{-1} \right) (\beta_* - (\phi_1 - \beta_*)) \right\}$$

其中 $\beta_* = (X_*^T X_* + \sigma^2 \phi_2^{-1})^{-1} X_*^T X_* (\phi_1 - \beta)$, 可推导出 β_* 的后验分布:

$$\beta_* | Y, b, a, \sigma^2 \sim N_{p+2} \left((\phi_1 - \beta_*), (\sigma^{-2} X_*^T X_* + \phi_2^{-1})^{-1} \right) \quad (5)$$

(2) 为随机效应 b_i 指定先验分布为 $b_i \sim N_q(0, D)$, 令 $w_i = Y_i - X_i \beta_* - Z^{(2)} a$, 根据似然表达式(4), 通过多元配方法, 推导出第 i 个个体的随机效应 b_i 的后验分布:

$$\pi(b_i | Y_i, \beta, D, \sigma^2) \propto \exp \left\{ -\frac{1}{2\sigma^2} (b_i - \hat{b}_i)^T (Z_i^{(1)})^T Z_i^{(1)} (b_i - \hat{b}_i) \right\} \exp \left\{ -\frac{1}{2} b_i^T D^{-1} b_i \right\} \\ \propto \exp \left\{ -\frac{1}{2} (b_i - b_i^*)^T \left(\frac{1}{\sigma^2} (Z_i^{(1)})^T Z_i^{(1)} + D^{-1} \right) (b_i - b_i^*) \right\}$$

其中, $\hat{b}_i = \left((Z_i^{(1)})^T Z_i^{(1)} \right)^{-1} (Z_i^{(1)})^T w_i$, $b_i^* = \left((Z_i^{(1)})^T Z_i^{(1)} + \sigma^2 D^{-1} \right)^{-1} (Z_i^{(1)})^T (Y_i - X_i \beta - Z^{(2)} a)$, 可推导出随机效应 b_i 的后验分布:

$$b_i | Y_i, \beta, D, \sigma^2 \sim N_q \left(b_i^*, \left(\sigma^{-2} (Z_i^{(1)})^T Z_i^{(1)} + D^{-1} \right)^{-1} \right) \quad (6)$$

(3) 为随机效应 a 指定先验分布为 $a \sim N_{r-2}(0, \tau I)$, 根据似然表达式(4), 通过多元配方法, 推导出 a 的条件后验分布:

$$\pi(a | Y, \beta_*, b, D, \tau, \sigma^2) \propto \exp \left\{ -\frac{1}{2} (a - a^*)^T \left(\frac{1}{\sigma^2} (Z^{(2)})^T Z^{(2)} + (\tau I)^{-1} \right) (a - a^*) \right\}$$

其中 $a^* = \left((Z^{(2)})^T Z^{(2)} + \sigma^2 (\tau I)^{-1} \right)^{-1} (Z^{(2)})^T (Y - X_* \beta_* - Z^{(1)} b)$, 可推导出随机效应 a 的条件后验分布:

$$a | Y, \beta_*, b, D, \tau, \sigma^2 \sim N_{r-2} \left(a^*, \left(\sigma^{-2} (Z^{(2)})^T Z^{(2)} + \tau^{-1} I \right)^{-1} \right) \quad (7)$$

(4) 为随机效应方差 D 指定先验分布为 $D \sim IW_q(o_o, R_o)$, 根据似然表达式(4), 通过多元配方法, 推导出 D 的条件后验分布:

$$\pi(D | b) \propto |D|^{-\frac{m}{2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^m b_i^T D^{-1} b_i \right\} |D|^{\frac{o_0+q+1}{2}} \exp \left\{ -\frac{1}{2} \text{tr}(D^{-1} R_o) \right\} \\ \propto |D|^{\frac{m+o_0+q+1}{2}} \exp \left\{ -\frac{1}{2} \text{tr} \left(D^{-1} \left(R_o + \sum_{i=1}^m b_i b_i^T \right) \right) \right\}$$

则随机效应方差 D 的条件后验分布:

$$D|b_i \sim IW_q\left(m+o_0, R_o + \sum_{i=1}^m b_i b_i^T\right) \quad (8)$$

(5) 为 τ 指定先验分布为 $\tau \sim IW(o_1, R_1)$, 根据似然表达式(4), 通过多元配方法, 推导出 τ 的条件后验分布:

$$\pi(\tau|a) \propto |\tau|^{\frac{r-2+o_1+1}{2}} \exp\left\{-\frac{1}{2}tr\left(\tau^{-1}(R_1 + a^T a)\right)\right\}$$

可得到 τ 的条件后验分布:

$$\tau|a \sim IW(r-2+o_1, R_1 + a^T a) \quad (9)$$

(6) 为 σ^2 指定先验分布 $\sigma^2 \sim 1/\sigma^2$, 根据似然表达式(4), 推导出 σ^2 的条件后验分布:

$$\pi(\sigma^2|\beta_*, b, a, Y) \propto (\sigma^2)^{-\left(\frac{n}{2}+1\right)} \exp\left\{-\frac{1}{\sigma^2}\left(\frac{1}{2}u^T u\right)\right\}$$

其中 $u = Y - X_*\beta_* - Z^{(1)}b - Z^{(2)}a$ 。

上述结果为逆 Gamma 分布, 则推导出 σ^2 的条件后验分布:

$$\sigma^2|Y, \beta_*, b, a \sim IG\left(\frac{n}{2}, \frac{1}{2}u^T u\right) \quad (10)$$

3.3. 贝叶斯估计

式(5)~(10)中参数的后验分布为标准形式, 可采用 Gibbs 抽样, Gibbs 抽样是一种马尔可夫链蒙特卡罗(MCMC)方法, 用于从多维概率分布中生成样本, 具体步骤如下:

- (1) 初始化: 选择初始值 $\theta^{(0)}$ 作为参数的起始点。
- (2) 迭代抽样: 对于每个参数 θ_i (在 $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ 中), 执行以下步骤:
 - a) 从条件分布 $p(\theta_i|\theta_{-i})$ 中抽样, 其中 θ_{-i} 表示除 θ_i 以外的所有参数;
 - b) 更新 θ_i 的值为新抽样的结果。
- (3) 重复迭代: 重复步骤(2), 直至达到预定的迭代次数或收敛标准。

设 $\{\beta^{(k)}, D^{(k)}, \tau^{(k)}, \sigma^{2(k)}: k=1, \dots, K\}$ 是由 Gibbs 抽样抽取的随机样本, 则它们的贝叶斯估计可以表示如下:

$$\hat{\beta} = \frac{1}{K} \sum_{k=1}^K \beta^{(k)}, \hat{\sigma}^2 = \frac{1}{K} \sum_{k=1}^K \sigma^{2(k)}, \hat{D} = \frac{1}{K} \sum_{k=1}^K D^{(k)}, \hat{\tau} = \frac{1}{K} \sum_{k=1}^K \tau^{(k)}$$

4. 实证分析

为了验证所提出方法的有效性, 本节将基于所提出的半参数混合效应模型对饮食类非遗产品网络营销的影响因素进行分析。在 166 个国家级饮食类非物质文化遗产项目中, 选择了在淘宝设有官方旗舰店的样本, 总计 80 个, 共进行了 12 次统计(2023 年 1 月~12 月)[13]研究, 定义网络销量为响应变量 Y , 并选取淘宝官方旗舰店近半年综合销量排名第一的产品的月销量。以网络搜索量(X_1)、短视频传播效果(X_2)、店铺经营年限(X_3)、店铺粉丝数量(X_4)和店铺好评率(X_5)作为解释变量建立如下半参数混合效应模型:

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} + \beta_5 X_{5ij} + f(t_{ij}) + b_i + \varepsilon_{ij} \quad (11)$$

其中， $i=1,\cdots,80$ ， $j=1,\cdots,12$ ， $f(t)$ 是一个以 t 为自变量的函数， b_i 是随机截距，服从正态分布 $N(0,D)$ 且相互独立， ε_{ij} 为误差项，假定其服从正态分布 $N(0,\sigma^2)$ 。

将数据代入模型(11)，通过贝叶斯方法估计出模型中的参数，估计结果见表 1。

Table 1. Parameter estimation results
表 1. 参数估计结果

参数	估计值	标准差
β_0	24.833	0.004
β_1	0.652	0.142
β_2	0.877	0.084
β_3	0.538	0.008
β_4	1.083	0.001
β_5	1.314	0.001
D	0.217	0.023
σ^2	0.209	0.019

从表 1 可以看出，截距项为 24.833，网络搜索量系数为 0.652，代表网络搜索量对于店铺的网络营销量有一个正向影响，即网络搜索量越高，店铺营销量越高。短视频传播效果系数为 0.877，代表短视频传播效果对于网络营销量有一个较强的正向影响，即短视频传播效果越好，营销量越高。店铺经营年限系数为 0.538，代表店铺经营年限对于店铺的网络营销量有一个正向影响，即店铺经营年限越长，网络营销量越高。店铺粉丝数量系数为 1.083，代表店铺粉丝量对店铺的网络营销量有一个较强的正向影响，即店铺粉丝量越多，店铺的网络营销量越高。店铺好评率系数为 1.314，说明店铺好评率对店铺的网络营销量有一个很强的正向影响，即店铺好评率越高，店铺的网络营销量越高。以上情况数据合理地解释了网络营销中的部分营销因素，与实际情况大致相符。

5. 结束语

本文构建了一个纵向数据的半参数混合效应模型，并运用贝叶斯方法进行了统计推断。研究发现，模型中的固定效应、随机效应及其方差，以及在先验分布下的随机误差的后验分布均符合标准分布，从而有效简化了模型估计的统计计算。同时，所提出的模型被应用于分析网络营销中的营销因素，实证结果显示这些因素对淘宝店铺的网络销售量产生了不同程度的影响。实证分析结论也与网络营销的实际情况相符，说明了模型的可行性和有效性。但是在实际情况中，店铺的网络营销量不仅受到文中所述的解释变量影响，还受到诸多因素的共同影响，如：支付方式、物流与配送、品牌信誉、市场营销策略等。如果忽视这些因素，可能会导致模型的准确性下降。因此，若能对这些因素进行量化并将其纳入模型进行计算，结果将会更加精准。这也是本文未来研究的一个重要方向。

参考文献

[1] 高源，张桂刚. 基于大数据的网络营销对策研究[J]. 湖北经济学院学报(人文社会科学版), 2014(2): 66-68.

- [2] 王明. 大数据时代网络营销策略分析[J]. 海峡科技与产业, 2015(12): 94-96.
- [3] 肖阵. 大数据时代电子商务网络营销探讨[J]. 现代商贸工业, 2015, 36(14): 50-51.
- [4] 张晓玲. 搜索引擎营销: 企业营销的新亮点[J]. 商业文化月刊, 2008(1): 252-253.
- [5] 许丽庆. 纵向数据模型的变量选择及其在网络营销中的应用[D]: [硕士学位论文]. 南宁: 广西大学, 2017.
- [6] Elmi, A., Ratcliffe, S.J., Parry, S. and Guo, W. (2011) A B-Spline Based Semiparametric Nonlinear Mixed Effects Model. *Journal of Computational and Graphical Statistics*, **20**, 492-509. <https://doi.org/10.1198/jcgs.2010.09001>
- [7] 阙焯, 黄振生. 部分线性混合效应模型的有效估计[J]. 应用概率统计, 2017, 33(5): 529-537.
- [8] Lindley, D.V. and Smith, A.F.M. (1972) Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **34**, 1-18. <https://doi.org/10.1111/j.2517-6161.1972.tb00885.x>
- [9] 齐培艳, 华文杰, 段西发. 半参数非线性混合效应模型的多重估算法[J]. 统计与决策, 2021, 37(24): 24-28.
- [10] Brown, E.R. and G. Ibrahim, J. (2003) A Bayesian Semiparametric Joint Hierarchical Model for Longitudinal and Survival Data. *Biometrics*, **59**, 221-228. <https://doi.org/10.1111/1541-0420.00028>
- [11] Rupper, D., Wand, M.P. and Carroll, R.J. (2003) *Semiparametric Regression*. Cambridge University Press.
- [12] Leonard, T. and Hsu, J.S.J. (2001) *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers*. Cambridge University Press, 5.
- [13] 刘旭玲, 刘佳婕, 刘世豪, 等. 饮食类非遗产品网络营销影响因素的组态及提升路径[J]. 美食研究, 2024, 41(2): 23-31.