

引用格式:耿俏,李志强,陈少东.海量数据下线性混合效应模型的估计算法[J].北京化工大学学报(自然科学版),2019,46(3):123-128.

GENG Qiao, LI ZhiQiang, CHEN ShaoDong. Estimation algorithm for a linear mixed effect model with massive data[J]. Journal of Beijing University of Chemical Technology (Natural Science), 2019,46(3):123-128.

海量数据下线性混合效应模型的估计算法

耿 俏 李志强* 陈少东

(北京化工大学 理学院, 北京 100029)

摘 要:基于以往文献提出线性混合效应模型参数的三步估计方法,避免了繁杂的极大似然估计迭代步骤。同时为进一步解决海量数据下计算估计时存在的存储瓶颈及计算时间过长问题,在海量纵向数据的两种不同数据格式下,分别基于三步估计方法利用分治算法计算模型参数的估计量。数值模拟和实证分析结果表明,本文所提出的三步估计方法和估计量的分治算法可以减轻计算负担,减少占用内存,解决内存不足的问题,并提高计算速度。

关键词:海量数据;纵向数据;线性混合效应模型;三步估计方法;分治算法

中图分类号: O212 **DOI:** 10.13543/j.bhxbzr.2019.03.019

引 言

纵向数据是通过在每个个体在不同的时间点上进行观测得到的数据集,广泛应用在经济学、医学等领域,具有组内数据相关、组间数据独立的特点。线性混合效应模型^[1-5]能够很好地刻画数据间的组内相关性和组间独立性,因此是分析处理纵向数据的常用模型之一。

随着信息、网络技术的迅速发展,数据增长变化速度惊人,呈现出数据海量趋势^[6]。传统的线性混合模型估计方法在海量数据下遇到一系列的挑战,如存储瓶颈、计算效率等问题^[7-8],因此有必要探求新的算法对以往线性混合效应模型的估计方法进行改进。

为了构造线性混合效应模型中系数的最优估计,Wu等^[5]提出了基于协方差结构的加权最小二乘法,但该估计方法的效率与纵向数据的协方差结构估计方法密切相关。常用的协方差结构估计方法有极大似然法和限制极大似然法等,然而这些方法需要进行反复的迭代计算及优化步骤,当数据量非常大时会导致很高的计算时间和计算量成本。为了

解决此类问题,Sun等^[9]在研究随机效应变系数模型时提出了一种方差分量估计的简便方法,该方法不需要进行迭代计算,能够适应海量数据情形。

然而直接利用矩阵形式的加权最小二乘法估计模型系数会遇到存储问题,因此分治算法被提出并在海量数据的统计计算中得到了广泛应用^[10-13]。分治算法的核心思想是先把复杂问题分解成几个子问题,找到这几个子问题的解法后,再利用合适的方法把每个子问题的结果组合起来得到整个问题的结果。Chang等^[11]提出的分治算法是解决海量数据分析问题的最实用方法;Guha等^[12]把最小二乘法与分治算法结合起来解决海量数据下线性回归模型的估计问题:首先将海量数据集分成一系列可管理的数据块,然后对每个数据块独立运行最小二乘方法,最后整合每块数据的结果得到最终的结果。但之前学者们也只是利用分治算法解决海量数据下简单数据模型的估计问题,对更复杂的、应用更广泛的线性混合效应模型在海量数据下的估计算法还并未研究。

本文基于文献[5]的线性混合效应模型系数的加权最小二乘估计方法和文献[9]的方差分量的估计方法以及分治算法,提出海量数据下线性混合效应模型参数的估计算法。首先提出线性混合模型参数的三步估计方法,避免了海量数据下通常所需的基于极大似然估计或限制极大似然估计的迭代计算,再针对海量数据的两种不同情形,基于分治算法

收稿日期:2018-09-15

第一作者:女,1993年生,硕士生

*通信联系人

E-mail: li-zhiqiang2000@163.com

分步骤计算估计量。实验表明本文算法不仅可以减少内存占用,解决存储问题,而且可以缩短计算时间,提高运算速度。

1 线性混合效应模型及其估计方法

1.1 线性混合效应模型

考虑纵向数据线性混合效应模型

$$Y_{ij} = X_{ij}^T \beta + Z_{ij}^T b_i + \varepsilon_{ij}, i = 1, \dots, n; j = 1, \dots, n_i \quad (1)$$

式中,未知参数 β 是 $p \times 1$ 维的固定效应向量, b_i 是 $q \times 1$ 维随机效应向量, X_{ij} 和 Z_{ij} 分别表示与之相关的协变量。假定随机效应 $b_i \sim N(0, \Sigma)$, 模型误差 $\varepsilon_{ij} \sim N(0, \sigma^2)$, 二者均满足独立同分布条件,且 b_i 与 ε_{ij} 相互独立。

1.2 三步估计方法

为了避免迭代计算,提出三步估计方法分别对线性混合效应模型的方差分量和固定效应进行估计,利用与文献[9]类似的证明步骤可以得到估计的相合性和渐近正态性,因此本文只考虑估计算法。

首先利用最小二乘法对系数进行初步估计。令 $Y = (Y_1^T, \dots, Y_n^T)^T$, $Y_i = (Y_{i1}, \dots, Y_{in_i})^T$, X 和 ε 类似定义, $b = (b_1^T, \dots, b_n^T)^T$, $Z_i = (Z_{i1}, \dots, Z_{in_i})^T$, $Z = \text{diag}(Z_1, \dots, Z_n)$, 则模型(1)可重新表示为

$$Y = X\beta + Zb + \varepsilon \quad (2)$$

对该模型系数直接求最小二乘估计,可得初步估计

$$\hat{\beta}_0 = (X^T X)^{-1} X^T Y \quad (3)$$

其次估计方差分量。利用文献[9]的原理来估计方差分量 σ^2, Σ 。

令 $U_{ij} = Z_{ij}^T b_i + \varepsilon_{ij}$, 则有 $Y_{ij} = X_{ij}^T \beta + U_{ij}$, 进一步有 $\hat{U}_{ij} = Y_{ij} - X_{ij}^T \hat{\beta}_0$ 。记 $U_i = (U_{i1}, \dots, U_{in_i})^T$, 则 $\hat{U}_i = (\hat{U}_{i1}, \dots, \hat{U}_{in_i})^T$, 进一步有 $U_i = Z_i b_i + \varepsilon_i$ 。从而利用最小二乘法进行估计可得随机效应 b_i 的估计值 $\tilde{b}_i = (Z_i^T Z_i)^{-1} Z_i^T U_i$, 其中 U_i 是未知的, 故用 U_i 的估计值 \hat{U}_i 代替 U_i 可得 $\hat{b}_i = (Z_i^T Z_i)^{-1} Z_i^T \hat{U}_i$ 。进一步记 $\tilde{U}_i = Z_i \tilde{b}_i = Z_i (Z_i^T Z_i)^{-1} Z_i^T U_i \equiv H_i U_i$, 则模型 $U_i = Z_i b_i + \varepsilon_i$ 的残差平方和为 $S_i = (U_i - \tilde{U}_i)^T (U_i - \tilde{U}_i) = U_i^T U_i - U_i^T H_i U_i$, 从而有

$$\hat{S}_i = \hat{U}_i^T \hat{U}_i - \hat{U}_i^T H_i \hat{U}_i \quad (4)$$

根据文献[9]的估计步骤,进一步考虑系数的估计值 $\hat{\beta}_0$ 的影响,可以类似构造出方差的估计

$$\hat{\sigma}^2 = \left[\sum_{i=1}^n (n_i - q) - p \right]^{-1} \sum_{i=1}^n \hat{S}_i = (N - qn -$$

$$p)^{-1} \sum_{i=1}^n \hat{S}_i \quad (5)$$

式中 $N = \sum_{i=1}^n n_i$, q 为 b_i 的维数, p 为 β 的维数。下一步估计 Σ 。根据 U_i 的定义可知

$$\tilde{b}_i = (Z_i^T Z_i)^{-1} Z_i^T U_i = b_i + (Z_i^T Z_i)^{-1} Z_i^T \varepsilon_i$$

进一步有

$$\begin{aligned} \sum_{i=1}^n \tilde{b}_i \tilde{b}_i^T &= \sum_{i=1}^n b_i b_i^T + \sum_{i=1}^n (Z_i^T Z_i)^{-1} Z_i^T \varepsilon_i \varepsilon_i^T Z_i \\ &\quad (Z_i^T Z_i)^{-1} + \sum_{i=1}^n (Z_i^T Z_i)^{-1} Z_i^T \varepsilon_i b_i^T + \sum_{i=1}^n b_i \varepsilon_i^T Z_i \\ &\quad (Z_i^T Z_i)^{-1} \end{aligned} \quad (6)$$

通过直接计算各项的一、二阶矩,可证式(6)中最后两项的阶为 $O_p(n^{-\frac{1}{2}})$, 因此相对于其他项可将其忽略不计,由此可得

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n b_i b_i^T &\approx \frac{1}{n} \left[\sum_{i=1}^n \tilde{b}_i \tilde{b}_i^T - \sum_{i=1}^n (Z_i^T Z_i)^{-1} \right. \\ &\quad \left. Z_i^T \varepsilon_i \varepsilon_i^T Z_i (Z_i^T Z_i)^{-1} \right] \approx \frac{1}{n} \left[\sum_{i=1}^n \tilde{b}_i \tilde{b}_i^T - \sigma^2 \sum_{i=1}^n (Z_i^T Z_i)^{-1} \right] \end{aligned}$$

从而可得

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \hat{b}_i \hat{b}_i^T - \frac{1}{n} \hat{\sigma}^2 \sum_{i=1}^n (Z_i^T Z_i)^{-1} \quad (7)$$

最后计算 β 的加权最小二乘估计,记

$$V = \text{diag}(V_1, \dots, V_n)$$

$$V_i = \text{var}(Z_i b_i + \varepsilon_i) = Z_i \Sigma Z_i^T + \sigma^2 I_{n_i}$$

由公式(6)、(7)有

$$\hat{V}_i = Z_i \hat{\Sigma} Z_i^T + \hat{\sigma}^2 I_{n_i}, \quad \hat{V} = \text{diag}(\hat{V}_1, \dots, \hat{V}_n)$$

根据文献[5]的加权最小二乘估计方法可得

$$\hat{\beta} = (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} Y \quad (8)$$

或利用文献[5]的样本求和公式(9)计算

$$\hat{\beta} = \left(\sum_{i=1}^n X_i^T \hat{V}_i^{-1} X_i \right)^{-1} \left(\sum_{i=1}^n X_i^T \hat{V}_i^{-1} Y_i \right) \quad (9)$$

2 海量数据下线性混合效应模型的分治算法

考虑海量数据的两种情形:①个体数据量较大,但组内数据量较小;②个体数据量较小,但组内数据量较大。此时会遇到存储瓶颈及计算低效的问题,1.2节中的估计方法不再适用。因此在此两种情形下分别考虑利用分治算法对前面提出的三步估计方法进行调整。

2.1 个体数据量较大, 组内数据量较小情形

令 $D = \{(\mathbf{X}_{ij}, \mathbf{Z}_{ij}, Y_{ij})\}_{i=1}^n \{j=1}^{n_i}$ 为全部数据集。根据分治算法, 将数据集 D 划分为 K 个子集 D_1, D_2, \dots, D_K , 每个子集的样本数 $m = n/K$, 每个子集包含的数据为 $\mathbf{X}_{kij}, \mathbf{Z}_{kij}, Y_{kij}$, 其中 $1 \leq j \leq n_{ki}, 1 \leq i \leq m, 1 \leq k \leq K$ 且每个子集包含的数据不能超过单机处理能力。

首先计算初步估计。

记 $\mathbf{X}_{ki} = (\mathbf{X}_{ki1}, \dots, \mathbf{X}_{kin_i})^T, \mathbf{X}_k = (\mathbf{X}_{k1}^T, \dots, \mathbf{X}_{km}^T)^T, \mathbf{Y}_k$ 和 $\boldsymbol{\varepsilon}_k$ 类似定义, $\mathbf{Z}_{ki} = (\mathbf{Z}_{ki1}, \dots, \mathbf{Z}_{kin_i})^T, \mathbf{Z}_k = \text{diag}(\mathbf{Z}_{k1}, \dots, \mathbf{Z}_{km}), \mathbf{b}_k = (\mathbf{b}_{k1}^T, \dots, \mathbf{b}_{km}^T)^T$ 。

则第 k 个子集对应的数据模型可写作

$$\mathbf{Y}_k = \mathbf{X}_k \boldsymbol{\beta} + \mathbf{Z}_k \mathbf{b}_k + \boldsymbol{\varepsilon}_k \quad (10)$$

对每个子数据集直接作最小二乘估计有

$\hat{\boldsymbol{\beta}}_{0k} = (\mathbf{X}_k^T \mathbf{X}_k)^{-1} \mathbf{X}_k^T \mathbf{Y}_k, 1 \leq k \leq K$, 根据分治算法, 整合各个子集的结果, 保存每组 $\mathbf{X}_k^T \mathbf{X}_k, \hat{\boldsymbol{\beta}}_{0k}$, 最后求得初步估计

$$\hat{\boldsymbol{\beta}}_0 = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \left(\sum_{k=1}^K \mathbf{X}_k^T \mathbf{X}_k \right)^{-1} \sum_{k=1}^K \mathbf{X}_k^T \mathbf{X}_k \hat{\boldsymbol{\beta}}_{0k} \quad (11)$$

其次利用分治算法估计方差分量。记 $\hat{\mathbf{U}}_{ki} = (\hat{U}_{ki1}, \dots, \hat{U}_{kin_i})^T, \hat{\mathbf{U}}_k = (\hat{\mathbf{U}}_{k1}, \dots, \hat{\mathbf{U}}_{km})^T$ 与公式(4)类似, 可得

$$\hat{S}_k = \hat{\mathbf{U}}_k^T \hat{\mathbf{U}}_k - \hat{\mathbf{U}}_k^T \mathbf{H}_k \hat{\mathbf{U}}_k \quad (12)$$

所以有

$$\hat{\sigma}^2 = (N - qn - p)^{-1} \sum_{k=1}^K \hat{S}_k \quad (13)$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{k=1}^K \sum_{i=1}^m \hat{\mathbf{b}}_{ki} \hat{\mathbf{b}}_{ki}^T - \frac{1}{n} \hat{\sigma}^2 \sum_{k=1}^K \sum_{i=1}^m (\mathbf{Z}_{ki}^T \mathbf{Z}_{ki})^{-1} \quad (14)$$

由 1.2 节可知, 式中 $\hat{\mathbf{b}}_{ki} = (\mathbf{Z}_{ki}^T \mathbf{Z}_{ki})^{-1} \mathbf{Z}_{ki}^T \hat{\mathbf{U}}_{ki}$ 。

公式(11) ~ (13) 中, 对应每个数据子集需计算保存的数据为 $\{\hat{\mathbf{U}}_k^T \hat{\mathbf{U}}_k, \hat{\mathbf{U}}_k^T \mathbf{H}_k \hat{\mathbf{U}}_k, \mathbf{Z}_{ki}^T \mathbf{Z}_{ik}, \mathbf{Z}_{ki}^T \hat{\mathbf{U}}_{ki}\}$ 。

最后计算模型系数 $\boldsymbol{\beta}$ 的加权估计。根据方差分量的计算公式(12) ~ (13) 可知, 第 k 个子数据模型的协方差矩阵的估计为 $\hat{\mathbf{V}}_k = \mathbf{Z}_k \hat{\boldsymbol{\Sigma}} \mathbf{Z}_k^T + \hat{\sigma}^2 \mathbf{I}_{n_i \times m}$ 。对每个子集用加权最小二乘法进行估计有 $\hat{\boldsymbol{\beta}}_k = (\mathbf{X}_k^T \hat{\mathbf{V}}_k^{-1} \mathbf{X}_k)^{-1} \mathbf{X}_k^T \hat{\mathbf{V}}_k^{-1} \mathbf{Y}_k$, 由此可知, 保留数据 $\mathbf{X}_k^T \hat{\mathbf{V}}_k^{-1} \mathbf{X}_k, \hat{\boldsymbol{\beta}}_k$, 把每个子集的结果整合起来得到

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{Y} = \left(\sum_{k=1}^K \mathbf{X}_k^T \hat{\mathbf{V}}_k^{-1} \mathbf{X}_k \right)^{-1} \sum_{k=1}^K \mathbf{X}_k^T \hat{\mathbf{V}}_k^{-1} \mathbf{X}_k \hat{\boldsymbol{\beta}}_k \quad (15)$$

2.2 个体数据量较小, 组内数据量较大情形

对每个个体运用分治算法。

令 $D_i = \{(\mathbf{X}_{ij}, \mathbf{Z}_{ij}, Y_{ij})\}_{j=1}^{n_i}, i = 1, \dots, n$, 将数据 D_i 划分为 K 个子集 $D_{i1}, D_{i2}, \dots, D_{iK}$, 每个子集的样本数为 $m_i = n_i/K$ 。每个子集包含的数据为 $\mathbf{X}_{ikj}, \mathbf{Z}_{ikj}, Y_{ikj}$, 其中 $1 \leq j \leq m_i, 1 \leq k \leq K, 1 \leq i \leq n$, 并且每个子集包含的数据不能超过单机处理能力。

首先计算初步估计。记 $\mathbf{X}_{ik} = (\mathbf{X}_{ik1}, \dots, \mathbf{X}_{ikm_i})^T, \mathbf{X}_i = (\mathbf{X}_{i1}^T, \dots, \mathbf{X}_{iK}^T)^T, \mathbf{Y}_i, \mathbf{Z}_i, \boldsymbol{\varepsilon}_i$ 类似定义, 第 i 个个体的第 k 个子集对应的数据模型为

$$\mathbf{Y}_{ik} = \mathbf{X}_{ik} \boldsymbol{\beta} + \mathbf{Z}_{ik} \mathbf{b}_i + \boldsymbol{\varepsilon}_{ik} \quad (16)$$

对每个个体的每个数据集直接作最小二乘估计有 $\hat{\boldsymbol{\beta}}_{0ik} = (\mathbf{X}_{ik}^T \mathbf{X}_{ik})^{-1} \mathbf{X}_{ik}^T \mathbf{Y}_{ik}, 1 \leq i \leq n, 1 \leq k \leq K$ 。

根据分治算法, 整合每个个体每个子集的结果, 保存每组 $\mathbf{X}_{ik}^T \mathbf{X}_{ik}, \hat{\boldsymbol{\beta}}_{0ik}$, 最后求出初步估计

$$\hat{\boldsymbol{\beta}}_0 = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \left[\sum_{i=1}^n \left(\sum_{k=1}^K \mathbf{X}_{ik}^T \mathbf{X}_{ik} \right)^{-1} \sum_{i=1}^n \left(\sum_{k=1}^K \mathbf{X}_{ik}^T \mathbf{X}_{ik} \hat{\boldsymbol{\beta}}_{0ik} \right) \right] \quad (17)$$

其次利用分治算法估计方差分量。记 $\hat{\mathbf{U}}_{ik} = (\hat{U}_{ik1}, \dots, \hat{U}_{ikm_i})^T$, 与公式(4)类似, 可得

$$\hat{S}_i = \sum_{k=1}^K \hat{\mathbf{U}}_{ik}^T \hat{\mathbf{U}}_{ik} - \sum_{k=1}^K \hat{\mathbf{U}}_{ik}^T \mathbf{H}_{ik} \hat{\mathbf{U}}_{ik} \quad (18)$$

所以有

$$\hat{\sigma}^2 = (N - qn - p)^{-1} \sum_{i=1}^n \hat{S}_i \quad (19)$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^n \hat{\mathbf{b}}_i \hat{\mathbf{b}}_i^T - \frac{1}{n} \hat{\sigma}^2 \sum_{i=1}^n \left(\sum_{k=1}^K \mathbf{Z}_{ik}^T \mathbf{Z}_{ik} \right)^{-1} \quad (20)$$

由 1.2 节可知, 式中 $\hat{\mathbf{b}}_i = \left(\sum_{k=1}^K \mathbf{Z}_{ik}^T \mathbf{Z}_{ik} \right)^{-1} \sum_{k=1}^K \mathbf{Z}_{ik}^T \hat{\mathbf{U}}_{ik}$ 。

根据公式(17) ~ (18) 可知, 对应每个个体子集, 需计算保存的数据有 $\{\hat{\mathbf{U}}_{ik}^T \hat{\mathbf{U}}_{ik}, \hat{\mathbf{U}}_{ik}^T \mathbf{H}_{ik} \hat{\mathbf{U}}_{ik}, \mathbf{Z}_{ik}^T \mathbf{Z}_{ik}, \mathbf{Z}_{ik}^T \hat{\mathbf{U}}_{ik}\}$ 。

最后计算模型系数 $\boldsymbol{\beta}$ 的加权估计。根据方差分量的计算公式(18) ~ (19), 第 i 个个体的第 k 个子数据模型的协方差矩阵的估计为 $\hat{\mathbf{V}}_{ik} = \mathbf{Z}_{ik} \hat{\boldsymbol{\Sigma}} \mathbf{Z}_{ik}^T + \hat{\sigma}^2 \mathbf{I}_{m_i}$, 对每个个体的每个子集运用加权最小二乘估计有

$$\hat{\boldsymbol{\beta}}_{ik} = (\mathbf{X}_{ik}^T \hat{\mathbf{V}}_{ik}^{-1} \mathbf{X}_{ik})^{-1} \mathbf{X}_{ik}^T \hat{\mathbf{V}}_{ik}^{-1} \mathbf{Y}_{ik} \quad (21)$$

由此可知, 对每个个体的每个子集保留数据 $\mathbf{X}_{ik}^T \hat{\mathbf{V}}_{ik}^{-1} \mathbf{X}_{ik}, \hat{\boldsymbol{\beta}}_{ik}$, 得到

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{Y} = \left[\sum_{i=1}^n \left(\sum_{k=1}^K \mathbf{X}_{ik}^T \hat{\mathbf{V}}_{ik}^{-1} \mathbf{X}_{ik} \right)^{-1} \sum_{i=1}^n \left(\sum_{k=1}^K \mathbf{X}_{ik}^T \hat{\mathbf{V}}_{ik}^{-1} \mathbf{X}_{ik} \hat{\boldsymbol{\beta}}_{ik} \right) \right]$$

$$\mathbf{X}_{ik} \Big) \Big]^{-1} \sum_{i=1}^n \Big(\sum_{k=1}^K \mathbf{X}_{ik}^T \hat{\mathbf{V}}_{ik}^{-1} \mathbf{X}_{ik} \hat{\boldsymbol{\beta}}_{ik} \Big) \tag{22}$$

3 数值模拟

为了将本文估计算法与极大似然法^[5]进行对比,验证本文算法在计算时间上的优势,同时检验算法的可行性,采用了 Matlab 软件进行数值模拟。

3.1 模拟数据

设纵向数据线性混合效应模型如下

$$Y_{ij} = \mathbf{X}_{ij}^T \boldsymbol{\beta} + \mathbf{Z}_{ij}^T \mathbf{b}_i + \varepsilon_{ij}, i = 1, \cdots, n; j = 1, \cdots, n_i \tag{23}$$

分别考虑海量数据下的两种样本情形。

情形(1) 个体数量较大,但组内数据量较小时,考虑两种样本:① $n = 10\,000, n_i = 10$,总样本量 100 000;② $n = 100\,000, n_i = 10$,总样本量 1 000 000。

情形(2) 个体有有限个,但组内数据量较大时,样本为 $n = 10, n_i = 10\,000$ 。

另外参数部分系数为: $\boldsymbol{\beta} = (1, 2, -3, 1, -2)^T$; $\mathbf{X}_{ij} = (X_{ij1}, X_{ij2}, X_{ij3}, X_{ij4}, X_{ij5})^T$, 其中 $X_{ij1} \sim N(0, 1)$, $X_{ij2} \sim N(0, 2)$, $X_{ij3} \sim N(0, 1)$, $X_{ij4} \sim N(0, 2)$, $X_{ij5} \sim N(0, 1)$, 且 $X_{ij1}, X_{ij2}, X_{ij3}, X_{ij4}, X_{ij5}$ 相互独立; $\mathbf{Z}_{ij} = (Z_{ij1}, Z_{ij2})^T$, Z_{ij1} 和 Z_{ij2} 都独立同分布于 $N(0, 2)$; $\mathbf{b}_i = (b_{i1}, b_{i2})^T$, b_{i1} 和 b_{i2} 都独立同分布于 $N(0, 1)$; $\varepsilon_{ij} \sim N(0, \sigma^2)$, 其中 $\sigma^2 = 2$, 且 b_i 与 ε_i 相互独立。模拟次数 50 次。

3.2 情形(1)的模拟结果与分析

在情形(1)的两种样本情况下,分别用极大似然法^[5]和本文算法估计模型系数和方差分量,并将两种方法所用的计算时间及估计精度进行对比。

3.2.1 估计精度和计算效率

利用均方误差(MSE)衡量模拟估计值与真实值之间的偏差,对估计精度进行分析,MSE 值越小,估计精度越高。均方误差计算公式如下

$$\gamma_{\text{MSE}} = \frac{1}{m} \sum_{k=1}^m (x_{\text{obs},k} - x_{\text{real}})^2 \tag{24}$$

式(24)中, m 是模拟次数, x_{real} 为参数真值, $x_{\text{obs},k}$ 是第 k 次模拟得到的参数估计值。

通过模拟得到情形(1)中两种样本量下的计算时间和均方误差的值如表 1 所示。

由表 1 可知,极大似然法与本文算法估计出的模型参数的 MSE 值都较小,说明参数的估计效果较好,估计精确度较高。另外,与极大似然法相比,本文算法估计参数可以在几乎不降低估计精确度的同

表 1 情形(1)下两种算法的计算时间及估计精度比较
Table 1 Comparison of calculation time and estimated accuracy of two algorithms under case (1)

样本	t^{a}/s	$\text{MSE}_{(\beta)}^{\text{a)}$	$\text{MSE}_{(\sigma^2)}^{\text{a)}$	t^{b}/s	$\text{MSE}_{(\beta)}^{\text{b)}$	$\text{MSE}_{(\sigma^2)}^{\text{b)}$
①	10.62	0.000 2	0.000 9	2.05	0.000 3	0.475 4
②	106.54	0.000 07	0.000 6	20.37	0.000 09	0.356 1

a)—极大似然法;b)—本文算法。

时将运行效率提高 4 倍,说明该算法可以大幅减少计算时间,提高计算速度。

3.2.2 数据集块数对计算时间的影响

为了研究计算时间与子数据集块数之间的关系,以说明利用分治算法进行分块计算的必要性,分别将两种样本量下的数据分成不同的数据集块数,记录估计参数所用时间,结果如表 2、3 所示。

表 2 样本①计算时间与数据集块数的关系
Table 2 The relationship between the calculation time and the number of data sets in sample ①

K	t/s	K	t/s
1	—	1 000	2.43
50	59.27	2 000	2.05
100	17.77	5 000	2.26
200	6.78	10 000	2.63 *
400	3.43		

* 样本求和方法所用时间。

表 3 样本②计算时间与数据集块数的关系
Table 3 The relationship between the calculation time and the number of data sets in sample ②

K	t/s	K	t/s
1	—	5 000	29.60
200	33 281.12	20 000	20.37
500	652.56	50 000	21.59
1 000	179.01	100 000	25.71 *
2 000	65.07		

* 样本求和方法所用时间。

由表 2 和表 3 看出,当 $K = 1$ 时,直接利用加权最小二乘公式(8)估计模型系数,此时会超出内存而无法计算。样本①、②下,即 K 分别取 10 000 和 100 000 时,利用样本求和公式(9)估计模型系数。 K 取表中其他值时,按照本文算法进行计算,当 K 缓慢增加时,所用时间逐渐减少;当 K 取一个恰当的值,如 K 分别取样本①、②中 2 000 和 20 000 时,计算时间达到最小,并且与样本求和方式相比,运行速

度可提高 20% 以上。这是因为当 K 增加时,每个数据集块分到的样本数量同时减少,因此计算量和计算时间也随之减少;同时在单机处理能力有限且固定情况下,一定数量级的数据在合适的单机数目下可以达到最佳计算速度。而实际中的数据量更大,通常达百万级甚至千万级,此时本文算法的效果会更好。

综上说明,本文算法可以降低内存开销,减少计算时间,提高计算速度。

3.3 情形(2)的模拟结果与分析

情形(2)下,本文算法所用计算时间与数据集块数的关系如表 4 所示。 $K = 1$ 时利用样本求和公式(9)计算系数; K 取其他值时利用本文算法计算,可以看出时间随数据块数的变化趋势与情形(1)类似,且 $K = 200$ 时 $t = 1.38\text{ s}$,计算时间达到最小。同时利用极大似然法计算情形(2)模型参数,所用时间为 2 269.71 s,再次证明了本文算法的高效性。

表 4 情形(2)下计算时间与数据集块数的关系

Table 4 The relationship between the calculation time and the number of data sets in case (2)

K	t/s	K	t/s
1	1 049.77 *	200	1.38
10	17.23	500	2.21
100	1.58	1 000	3.83

* 样本求和方法所用时间。

4 实证分析

选取 2006—2010 年全国的 285 个地级及以上城市的地区生产总值、工业总产值和社会消费品零售总额、固定资产投资与城乡居民储蓄年末余额数据,实证研究工业总产值和社会消费品零售总额对地区生产总值的影响。

以地区生产总值(Y)作为响应变量,工业总产值(X_1)和社会消费品零售总额(X_2)作为固定效应的协变量;考虑到固定资产投资(Z_1)和城乡居民储蓄年末余额(Z_2)也会影响地区生产总值,故将这两个因素作为随机效应部分,建立线性混合效应模型

$$Y_{ij} = X_{ij}^T \boldsymbol{\beta} + Z_{ij}^T \boldsymbol{b}_i + \varepsilon_{ij}, i = 1, \cdots, 285; j = 1, \cdots, 5$$

(25)

将数据带入模型(25)可知,此时数据结构符合情形(1),因此利用情形(1)下的估计算法进行参数估计,具体结果见表 5。

表 5 $\boldsymbol{\beta}$ 的估计值和标准差

Table 5 Estimated value and standard deviation of $\boldsymbol{\beta}$

参数	β_1	β_2
估计值	0.223 1	1.792 9
标准差	0.010 0	0.037 4

由表 5 可知,工业总产值和社会消费品零售总额的系数都为正值,说明这两个因素正面影响地区生产总值,且所建模型是有效的。所以要提高地区生产总值,就要努力提高工业生产水平,增加社会消费品零售总额,提高民众收入,改善消费环境,稳定物价水平,解决医疗、卫生、就业等问题。

5 结束语

本文针对海量数据下的两种情形,将线性混合效应模型系数和方差分量的估计方法与分治算法相结合,提出了三步估计算法,使得复杂的统计模型也可以应用于海量数据情形。数值模拟表明该算法可以大幅减少计算时间,提高计算效率,并解决海量数据下计算机内存不足的问题。通过实证分析证明了本文模型及估计算法在实际应用中的可行性。

下一步,可以考虑采用分治算法与其他估计方法结合,从而实现其他统计模型在海量数据下的估计与应用。

参考文献:

[1] JIANG J. Linear and generalized linear mixed models and their applications [M]. New York: Springer, 2007.

[2] 李光辉. 线性混合模型中的参数估计[D]. 西宁: 青海师范大学, 2010.
LI G H. Parameter estimation in linear mixed model[D]. Xining: Qing Hai Normal University, 2010. (in Chinese)

[3] 许玉莉. 线性混合效应模型中方差分量的估计[J]. 应用概率统计, 2009, 25(3): 301–308.
XU W L. Estimation of variance components in linear mixed effects models[J]. Application Probability Statistics, 2009, 25(3): 301–308. (in Chinese)

[4] 孙燕, 柴根象. 纵向数据线性混合效应模型的统计分析[J]. 应用科学学报, 2006, 24(1): 70–73.
SUN Y, CHAI G X. Statistical analysis of linear mixed effect model for longitudinal data[J]. Journal of Applied Science, 2006, 24(1): 70–73. (in Chinese)

[5] WU H L, ZHANG J T. Nonparametric regression methods for longitudinal data analysis [M]. Hoboken: John

- Wiley & Sons, Inc., 2006: 17–25.
- [6] WU X D, ZHU X Q, WU G Q, et al. Data mining with big data[J]. IEEE Transactions on Knowledge & Data Engineering, 2013, 26(1): 97–107.
- [7] FAN J Q, HAN F, LIU H. Challenges of big data analysis[J]. National Science Review, 2014, 1(2): 293–314.
- [8] ZHOU Z H, CHAWLA N V, WILLIAMS G J, et al. Big data opportunities and challenges: discussions from data analytics perspectives [discussion forum] [J]. IEEE Computational Intelligence Magazine, 2014, 9(4): 62–74.
- [9] SUN Y, ZHANG W Y, TONG H. Estimation of the covariance matrix of random effects in longitudinal studies [J]. Annals of Statistics, 2007, 35(6): 2795–2814.
- [10] LIN N, XI R B. Aggregated estimating equation estimation[J]. Statistics & Its Interface, 2011, 1(1): 73–83.
- [11] CHANG X Y, LIN S B, WANG Y. Divide and conquer local average regression[J]. Electronic Journal of Statistics, 2017, 11(1): 1326–1350.
- [12] GUHA S, HAFEN R, ROUNDS J, et al. Large complex data: divide and recombine (d&r) with rhipe[J]. The ISI's Journal for the Rapid Dissemination of Statistics Research, 2015, 1(1): 53–67.
- [13] CHEN X Y, XIE M G. A split-and-conquer approach for analysis of extraordinarily large data[J]. Statistica Sinica, 2014, 24(4): 1655–1684.

Estimation algorithm for a linear mixed effect model with massive data

GENG Qiao LI ZhiQiang* CHEN ShaoDong

(Faculty of Science, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract: Based on the previous literature, a three-step method for the estimation of the parameters of a linear mixed effect model has been proposed, which avoids the complicated iterative steps of maximum likelihood estimation. At the same time, in order to further solve the storage bottleneck and calculation time when calculating the estimator with massive data, the estimator of the model parameters has been calculated using the divide-and-conquer algorithm based on the three-step estimation method for two different data formats of massive vertical data. The results of numerical simulation and empirical analysis show that the three-step estimation method and the estimator divide-and-conquer algorithm proposed in this paper can reduce the computational burden, reduce the memory consumption, solve the problem of insufficient memory, and improve the calculation speed.

Key words: massive data; longitudinal data; linear mixed effect model; three-step estimation method; divide-and-conquer algorithm

(责任编辑:汪 琴)