【理论与方法】

异质性大数据的分布式估计

郭婧璇　徐慧超　祝婉晴　田茂再

【摘　要】随着物联网技术的进步，大数据给网络带宽和计算机存储能力带来巨大挑战，传统的集中式数据处理难以实现，客观上促进了分布式统计学习的发展。在无迭代算法研究中，Zhang 等（2013）证明了当数据集个数 $s = O(\sqrt{N})$ 时，基于局部经验风险最小化的分治（DC）简单平均估计量具有 $O(N^{-1})$ 均方误差收敛速度，Huang 和 Huo（2019）在 M 估计框架下进一步提出分布式一步估计量，但上述方法均未考虑海量数据可能存在的异质性对分治估计效果的影响。本文在线性模型框架下提出海量异质数据的分治一步加权估计，证明了估计量的渐近性质并考虑了异质性检验问题。将本文提出的方法应用于美国医疗保险实际数据分析，结果表明该方法能更好地拟合数据的线性趋势且显著提高了计算效率。

【关 键 词】分治策略；一步估计；海量数据；异质性；医疗保险

【作者简介】郭婧璇，中国人民大学统计学院博士研究生，研究方向为大数据建模与应用、函数型数据分析和诊断；徐慧超，中国人民大学统计学院博士研究生，研究方向为保险统计和交通运输统计；祝婉晴，中国人民大学统计学院硕士研究生，研究方向为经济统计分析；田茂再（通讯作者），新疆财经大学“天山学者”特聘教授，兰州财经大学“兴隆学者”特聘教授，教育部人文社会科学重点研究基地中国人民大学应用统计科学研究中心副主任，中国人民大学教授，博士生导师，杰出学者，研究方向为复杂数据分析，电子邮箱：mztian@ruc.edu.cn。

【原文出处】《统计研究》（京），2020.10.104～114

【基金项目】中国人民大学科学研究基金（中央高校基本科研业务费专项资金资助）项目“大数据分析的稳健统计理论与应用研究”（18XNL012）。

一、引言

物联网技术的广泛应用引起了数据量爆发式增长，使基于集中式处理的云计算暴露出诸多缺陷：网络负担加重且传输效率变低，物联网服务实时响应和敏感信息的私密性要求难以保证。边缘计算是解决上述问题的有效手段，其核心思想是通过分布式处理实现智能前端化，将原本由云处理的大型服务分散到边缘节点进行处理。早在 20 世纪 90 年代，计算机科学家已开始利用个人计算机的闲置处理能力解决大型科学计算问题，如梅森素数搜索计划（GIMPS）、地外文明发现计划（SETI@HOME）等。数据科学家也在分布式框架下更新了统计和机器学习方法，如 EM 算法、支持向量机和主成分分析等。层出不穷的高效通信数值方法为统计和机器学习的应用开辟了更广阔的空间。

现有分布式统计最优化算法可分为两类：迭代算法和无迭代算法。迭代算法中具有代表性的是 Boyd 等（2010）提出的乘数交替方向法（ADMM）和 Shamir 等（2014）提出的分布式近似牛顿法（DANE）。迭代算法可达到几何收敛速度，在一定条件下甚至可以达到线性收敛速度，但数据量巨大、数据集众多时，迭代通信的成本和效率问题将

会凸显;而无迭代算法避免了多次迭代通信,以牺牲均方误差收敛速度提高计算效率。Zhang 等(2013)证明了局部经验风险函数简单平均估计量的均方误差收敛速度为 $O(N^{-1}+(N/s)^{-2})$,其中 N 为总数据量,s 为局部数据集个数,由于想法直接、操作简便,简单平均估计在分布式统计问题研究中应用广泛。Shi 等(2018)提出局部 M 估计的加权平均估计量,对于 $O(N^{-1/3})$ 局部 M 估计具有更快的全局收敛速度。Huang 和 Huo(2019)折中两类算法,对简单平均估计量做一步 Newton - Raphson 迭代,使之具有相对于迭代估计更低的通信成本和优于平均估计的收敛速度。上述研究成为分治海量数据统计分析框架的基础。

在海量数据背景下,由于数据收集、传输和管理的限制,完整数据集通常难以得到,统计推断必须借助局部数据集的推断结果。作为研究变量间关系的重要统计方法,海量数据分治回归分析具有极强的理论价值和现实意义。减小估计偏差是海量数据回归方法研究中的重要问题,如 Lian 等(2019)基于投影样条提出高维部分线性模型中非参函数的分治"减偏"估计量。Lin 和 Li(2019)借鉴统计复合方法给出惩罚回归有偏局部估计量的全局纠偏分治算法,为从统计学角度研究分治算法开辟了新思路。由于无迭代算法应用便利,现有分治回归方法大多基于平均估计,迭代算法和一步算法鲜有出现。Chen 等(2019)给出了分位回归的分治线性估计和相应的类牛顿迭代算法,成为现有研究中少见的迭代分治回归算法。

由于数据来源的多样性,局部数据集异质现象较为常见,若在聚合局部结果时忽视该问题,则全局结果也失去了意义,现有文献中对该问题的讨论较少。Zhao 等(2016)和 Wang 等(2017)基于部分线性和可加线性模型对同质部分和异质部分进行分别推断,但上述方法在聚合局部估计结果时都采取了无迭代的平均估计,无法克服海量异质性数据回归估计收敛速度慢的弱点。

分治策略下的假设检验是统计学的另一个重要问题。Battey 等(2018)考虑了高维稀疏线性和广义模型的分布式检验和估计问题,提出偏差减小的聚合估计量,并根据其渐近性质构造得分检验统计量;针对参数的异质性检验,非海量数据设定的现有文献大多关注异方差问题,如梅波和徐礼文(2016)提出异方差多正态总体均值的加权极大似然 Bootstrap 检验,苏鹏和田茂再(2018)考虑了复合分位回归模型尺度参数的估计并构造 von - Neumann 统计量检验参数的异质性;数据量很大时,子总体的位置参数也可能存在异质性,Lu 等(2016)基于惩罚似然理论分别研究了异质非参变系数模型的简单、成对和同时检验问题,建立起完整的异质系数检验体系,但其检验统计量仍基于局部风险最小化估计的简单平均聚合。综合来看,当前海量数据的异质性均值检验问题受到的关注较少,相关研究仍不充分。

此外,分治策略与样本选择、模型平均等经典的统计问题紧密相关,相关成果参见张新雨和邹国华(2011)、陈心洁等(2015),其中讨论了多种模型平均权重选择方法,如基于信息准则和渐近最优准则,这些方法有助于改进分治简单平均的聚合效果。

本文在线性模型框架下考虑了海量异质性数据的估计和检验问题,提出了分治一步加权估计并证明了其渐近正态性。文章剩余部分安排如下:第二部分提出了基于 plug - in 局部估计分治一步加权平均估计量和算法;第三部分给出了上述估计的渐近性质并考虑了异质系数的检验问题;第四部分为实证分析,利用上述模型与方法对美国医疗服务提供者的医疗保险使用和支付数据进行分析;最后一部分为结果讨论与总结。

二、参数估计

(一)模型设定

数据量为 N 的完整数据集 S 可划分为 s 个独立的局部数据集,$\mathcal{J}=\{1,\cdots,s\}$。数据集 $S_j, j\in\mathcal{J}$ 上因变量 $y\in\mathbb{R}$ 关于异质变量 $x=(x_1,\cdots,x_d)^T\in\mathbb{R}^d$ 和同质变量 $z=(z_1,\cdots,z_k)^T\in\mathbb{R}^k$ 有如下关系:

$$y = x^T\beta_j + z^T\gamma + \epsilon \tag{1}$$

其中，$\beta_j = (\beta_{1j}, \cdots, \beta_{dj})^T$ 和 $\gamma = (\gamma_1, \cdots, \gamma_k)^T$ 为对应异质和同质系数向量；d 为异质变量的个数，k 为同质变量的个数。ϵ 均值为 0，方差为 σ^2。

为估计系数 γ 和 $\beta_j, j \in \mathcal{J}$，建立局部数据集 $S_j = \{(y_{ij}, x_{ij}, z_{ij})\}_{i=1}^{n_j}$ 的多元线性回归模型：

$$Y_j = X_j\beta_j + Z_j\gamma + e_j \tag{2}$$

其中，$Y_j = (y_{1j}, \cdots, y_{n_j j})^T$ 和 $e_j = (e_{1j}, \cdots, e_{n_j j})^T$ 为 n_j 维向量，$X_j = (x_{1j}, \cdots, x_{n_j j})^T$ 和 $Z_j = (z_{1j}, \cdots, z_{n_j j})^T$ 为设计矩阵。对于上述估计问题，数据集 S_j 的局部经验判别函数为：

$$M_j(\beta_j, \gamma) = \frac{1}{n_j}\sum_{i=1}^{n_j} m(\beta_j, \gamma) \tag{3}$$

其中，$m(\beta_j, \gamma) := (y, x, z; \beta_j, \gamma)$ 为 S_j 上的损失函数，本文取 $m(\beta_j, \gamma) = 1/2(y - x^T\beta_j - z^T\gamma)^2$。

完整数据集 $S = \sum_{j=1}^{s} S_j$ 的全局经验风险函数 $M(\beta_1, \cdots, \beta_s, \gamma)$ 为：

$$M(\beta_1, \cdots, \beta_s, \gamma) = \frac{1}{s}\sum_{j=1}^{s} M_j(\beta_j, \gamma) \tag{4}$$

数据集 S_j 上系数的总体风险函数最小化估计定义为：

$$(\beta_{0j}, \gamma_{0j}) = \operatorname{argmin}_{\beta_j \in \mathbb{R}^d, \gamma \in \mathbb{R}^k} \{M_{0j}(\beta_j, \gamma) = E[m(\beta_j, \gamma)]\} \tag{5}$$

本文涉及的向量范数为 $\|\cdot\|_2$ 和 $\|\cdot\|_\infty$，对于 $v = (v_1, \cdots, v_n) \in \mathbb{R}^n$，$\|v\|_\infty = \max_{1 \leqslant i \leqslant n} |v_i|$；矩阵范数为 $\|\,|\cdot|\,\|$，对于 $A \in \mathbb{R}^{m \times m}$，有 $\|\,|A|\,\| = \sup_{\mu:\mu \in \mathbb{R}^m, \|\mu\| \leqslant 1} \|A\mu\|_2$，为矩阵的最大奇异值。

（二）基于 plug－in 局部估计的分治一步估计

局部数据集 $S_j, j \in J$ 上系数 (β_j, γ) 的估计问题如下：

$$(\hat{\beta}, \hat{\gamma}) = \operatorname{argmin}_{\beta_j \in \mathbb{R}^d, \gamma \in \mathbb{R}^k} \frac{1}{2}\|Y_j - X_j\beta_j - Z_j\gamma\|_2^2 \tag{6}$$

若给定同质系数 γ，则有 β_j 的显式估计：

$$\hat{\beta}(\gamma) = (X_j^T X_j)^{-1} X_j^T (Y_j - Z_j\gamma) \tag{7}$$

γ 的 plug－in 局部估计量为：

$$\hat{\gamma}_j = (Z_j^T(I - P_j)Z_j)^{-1} Z_j^T (I - P_j) Y_j \tag{8}$$

同质系数 γ 的分治简单平均初始估计为：

$$\hat{\gamma} = \frac{1}{s}\sum_{j=1}^{s} \hat{\gamma}_j \tag{9}$$

系数 γ 和 $\beta_j, j \in \mathcal{J}$ 的分治一步估计分别为：

$$\tilde{\gamma} = \hat{\gamma} - [\ddot{M}(\hat{\gamma})]^{-1} \dot{M}(\hat{\gamma}) \tag{10}$$

$$\tilde{\beta}_j := \hat{\beta}(\tilde{\gamma}) = (X_j^T X_j)^{-1} X_j^T (Y_j - Z_j\tilde{\gamma}) \tag{11}$$

$\dot{M}(\hat{\gamma})$ 和 $\ddot{M}(\hat{\gamma})$ 分别为 γ 的全局经验风险函数 $M(\gamma) = M(\hat{\beta}_1(\gamma), \cdots, \hat{\beta}_s(\gamma), \gamma)$ 的梯度向量和海塞矩阵在 $\gamma = \hat{\gamma}$ 处的取值。

γ 基于 plug－in 局部估计的全局经验风险函数为：

$$M(\gamma) = \frac{1}{2s}\sum_{j=1}^{s} \|(I - P_j)(Y_j - Z_j\gamma)\|_2^2 \tag{12}$$

其中，$P_j = X_j(X_j^T X_j)^{-1} X_j^T$ 为投影矩阵，有 $\hat{E}(Z_j \mid X_j) = P_j Z_j$。对应的梯度向量为：

$$\dot{M}(\gamma) = -\frac{1}{s}\sum_{j=1}^{s} Z_j^T (I - P_j)(Y_j - Z_j\gamma) \tag{13}$$

海塞矩阵为：

$$\ddot{M}(\gamma) = \frac{1}{s}\sum_{j=1}^{s} Z_j^T (I - P_j) Z_j \tag{14}$$

局部数据集规模差异过大会降低估计量的均方误差收敛速度，因此在应用中可考虑将加权平均估计量作为初始估计，即：

$$\hat{\gamma}_w = \sum_{j=1}^{s} \frac{n_j}{N}\hat{\gamma}_j \tag{15}$$

（三）分治一步加权估计算法

输入：局部数据集 $S_j, j = 1, \cdots, s$；

输出：系数估计值 $\tilde{\gamma}$ 和 $\tilde{\beta}_j, j = 1, \cdots, s$；

1 $\ddot{M}(\hat{\gamma}) = 0, M(\hat{\gamma}) = 0$；

2 $\hat{\gamma} = \hat{\gamma}_1 = (Z_1^T(I - P_1)Z_1)^{-1} Z_1^T (I - P_1) Y_1$，$n = n_1$；

3 for $j = 2, \cdots, s$; do

4 $\quad \hat{\gamma}_j = (Z_j^T(I - P_j)Z_j)^{-1} Z_j^T (I - P_j) Y_j$；

5 $\quad \hat{\gamma} = (n + n_j)^{-1}(n\hat{\gamma} + n_j\hat{\gamma}_j)$，$n = n + n_j$；

6 for $j = 1, \cdots, s$; *do*

7 $\dot{M}_j(\hat{\gamma}) = -Z_j^T(I-P_j)(Y_j - Z_j\hat{\gamma})$, $\ddot{M}_j(\hat{\gamma}) = Z_j^T(I-P_j)Z_j$;

8 $\dot{M}(\hat{\gamma}) = (1-j^{-1})\dot{M}(\hat{\gamma}) + j^{-1}\dot{M}_j(\hat{\gamma})$, $\ddot{M}(\hat{\gamma}) = (1-j^{-1})\ddot{M}(\hat{\gamma}) + j^{-1}\ddot{M}_j(\hat{\gamma})$;

9 $\tilde{\gamma} = \hat{\gamma} - [\ddot{M}(\hat{\gamma})]^{-1}\dot{M}(\hat{\gamma})$;

10 for $j = 1, \cdots, s$; do

11 $\tilde{\beta}_j = (X_j^T X_j)^{-1} X_j^T (Y_j - Z_j\tilde{\gamma})$;

三、假设检验

(一)渐近性质

1. 主要假定。考察分治一步估计渐近性质的正则条件:

(1)对于局部数据集 S_j 上待估系数 (β_j, γ),其参数空间 $\Theta_j = \mathbb{R}^{d+k}$ 为紧凸集。

(2)记 $u = (x^T, z^T)^T$, $\theta_j = (\beta_j, \gamma)$,有海塞矩阵 $\ddot{M}_{0j}(\theta_j) = E(uu^T)$ 非负定。

(3)损失函数 $m(\theta_j)$ 具有二阶连续导数,记 $\theta_{0j} = (\beta_{0j}, \gamma_{0j})$; $\forall \theta_j \in B_{\delta_j}$,存在非零常数 G 和 H,

$$E[\|\dot{m}(\theta_j)\|_2^8] \leqslant G^8, E[\|\ddot{m}(\theta_j) - \ddot{M}_{0j}(\theta_j)\|^8] \leqslant H^8 \tag{16}$$

对 $\forall u \in \mathbb{R}^{d+k}$ 都成立,其中 $B_{\delta_j} = \{\theta_j \in \Theta_j : \|\theta_j - \theta_{0j}\|_2 \leqslant \delta_j\}$, δ_j 为非零常数; $\dot{m}(\theta_j) = uu^T$, $j \in J$ 与 θ_j 的取值无关,满足 $L(x)$—李普希茨连续条件: $\forall \theta_j, \theta'_j \in B_{\delta_j}$,

$$\|\ddot{m}(\theta_j) - \ddot{m}(\theta'_j)\| \leqslant L(x)\|\theta_j - \theta'_j\|_2 \tag{17}$$

其中, $\exists L > 0$,使 $E[L(x)^8] \leqslant L^8$, $E[(L(x) - E[L(x)]^8)] \leqslant L^8$。

2. 渐近正态性。

定理1 若局部数据集个数 $s = O(\sqrt{N})$,则同质系数的分治一步估计 $\tilde{\gamma}$ 相合且渐近正态,对 $N \to \infty$ 有, $\tilde{\gamma} - \gamma \xrightarrow{p} 0$, $\sqrt{N}(\tilde{\gamma} - \gamma) \xrightarrow{d} N(0, \Sigma)$,其中, $\Sigma = \sigma^2[E(\tilde{z}\tilde{z}^T)]^{-1}$, $\tilde{z} = z - E(z|x)$。

定理2 若局部数据集个数 $s = O(\sqrt{N})$,则异质系数的分治一步估计 $\tilde{\beta}_j$ 相合且渐近正态,对 $N \to \infty$ 有, $\tilde{\beta}_j - \beta_j \xrightarrow{p} 0$, $\sqrt{n_j}(\tilde{\beta}_j - \beta_j) \xrightarrow{d} N(0, A)$,其中, $A = \sigma^2[E(xx^T)]^{-1}$。

(二)异质性检验

1. 简单检验。为简化讨论,考虑局部数据量 $n_1 = \cdots = n_s = n$。对于系数 $\beta_j, j \in J$ 考虑简单检验问题:

$H_0: Q\beta_j = Q\beta'_j, \forall j \in \mathcal{J}$ *vs.* $H_1: Q\beta_j \neq Q\beta'_j, \exists j \in \mathcal{J}$

其中, β'_j 为给定的 d 维向量, $Q = (q_1^T, \cdots, q_r^T)^T \in \mathbb{R}^{r \times d}$ 为给定矩阵, $r \leqslant d$。定义检验统计量:

$$T_s = \max_{j \in \mathcal{J}} \sqrt{n}\|Q(\tilde{\beta}_j - \beta'_j)\|_\infty \tag{18}$$

其 Bootstrap 近似值为,

$$W_s = \max_{j \in \mathcal{J}} \sqrt{n}\|Q(X_j^T X_j)^{-1} X_j^T e_j\|_\infty \tag{19}$$

其中, $e_j \sim N(0, \sigma^2 I_n)$。

2. 成对检验。异质系数的成对检验问题如下:

$H_0: \beta_p - \beta_q = 0, \forall (p,q) \in \mathcal{G}$ *vs.* $H_1: \beta_p - \beta_q \neq 0, \exists (p,q) \in \mathcal{G}$

其中, $\mathcal{G} = \{(p,q): p, q \in \sqrt{N}, p \neq q\}$。给定显著性水平 α,构造 *Wald* 检验统计量:

$$\phi = I\{\sqrt{n}(\tilde{\beta}_p - \tilde{\beta}_q) \leqslant \sqrt{2}\tilde{A}^{\frac{1}{2}} z_{1-\alpha/2}\} \tag{20}$$

其中, $\tilde{A} = (\frac{1}{n}\sum_{j=1}^s X_j^T X_j)^{-1} \sum_{j=1}^s \tilde{\sigma}_j^2$, $\tilde{\sigma}_j^2$ 为 σ^2 在局部数据集 S_j 上的样本矩估计。若考虑更一般的成对检验统计量,

$$T'_s = \max_{(p,q) \in g} \sqrt{n}\|Q(\tilde{\beta}_p - \tilde{\beta}_q)\|_\infty \tag{21}$$

Bootstrap 近似值为:

$$W_s = \max_g \sqrt{2n}\|\sum_g Q(\sum_g X_j^T X_j)^{-1} X_j^T e_j\|_\infty \tag{22}$$

3. 同时检验。对系数 $\beta_j \in \mathbb{R}^d, j \in \mathcal{J}$ 考虑如下同时检验问题:

$H_0: \alpha^T\beta_1 = \alpha^T\beta_2 = \cdots = \alpha^T\beta_s$ *vs.* $H_1: \alpha^T\beta_p \neq \alpha^T\beta_q, \exists (p,q) \in \mathcal{G}$

其中，$\alpha \in \mathbb{R}^d$ 为给定 d 维向量。$\alpha^T\beta_j$ 局部估计 $\alpha^T\breve{\beta}_j$ 的方差估计量：

$$r^2 = \frac{1}{s-1}\sum_{j=1}^{s}(\alpha^T\breve{\beta}_j - \alpha^T\bar{\beta})^2 \tag{23}$$

其中，$\breve{\beta}_j = \hat{\beta}_j(\hat{\gamma}_j)$，$\hat{\gamma}_j$ 为 γ 在数据集 S_j 上的局部估计，$j \in \mathcal{J}$；$\bar{\beta} = \frac{1}{s}\sum_{j=1}^{s}\breve{\beta}_j$。原假设 H_0 成立时，r^2 是估计量 $\alpha^T\breve{\beta}_j$ 方差的渐近无偏估计，相似的，

$$\delta^2 = \frac{1}{2(s-1)}\sum_{j=2}^{s}(\alpha^T\breve{\beta}_j - \alpha^T\breve{\beta}_{j-1})^2 \tag{24}$$

H_0 成立时，也是估计量 $\alpha^T\breve{\beta}_j$ 方差的渐近无偏估计。

与苏鹏和田茂再（2018）相似，构造 von - Neumann 检验统计量：

$$T = \frac{r^2}{\delta^2} = \frac{2\sum_{j=1}^{m}(\alpha^T\breve{\beta}_j - \alpha^T\bar{\beta})^2}{\sum_{j=2}^{m}(\alpha^T\breve{\beta}_j - \alpha^T\tilde{\beta}_{j-1})^2 d} \tag{25}$$

H_0 成立时，有 $\sqrt{s}(T-1) \xrightarrow{d} N(0,1)$，$N \to \infty$。给定的显著性水平 α，对应 Wald 检验统计量为：

$$\varphi = I\{\sqrt{s}(T-1) \notin z_{1-\alpha/2}\} \tag{26}$$

四、实证分析

（一）数据来源

美国医疗保险体系建立在自由市场制度之上，长期以来因费用高昂、制度复杂等问题饱受诟病，然而其不完善性也促进了相关制度的探索和试验。区别于我国的报销式医疗保险，管理式医疗是美国当前主流的医疗保险形式，基本特点是：保险机构通过签约形式向医疗服务提供者支付定额医药费，由其利用该费用向参保人提供整套综合性医疗服务，实现保险机构和医疗服务提供者风险共担。该模式对我国保险制度的健全和改革具有重要参考价值。本文运用分治一步估计算法分析美国医疗保险数据，建立标准化医疗保险平均支付金额（AMPamt）分析和预测模型。该数据来源于美国医疗保险和医疗补助服务中心（CMS）官网的“Medicare Physician and Other Supplier Data CY 2016”数据集，包含美国 51 个行政州和特区医疗服务提供者的近 1000 万条服务记录，涵盖了提供者的详细信息和服务细节。感兴趣分类变量如下：①医疗服务提供者属性，用于区分服务者为个人或组织；②提供者类型，包含流感中心等 90 余种具体分类，以特定编码进行区分；③医疗保险计划参与状况，即服务者是否接受保险允许金额分配方案；④服务地点，用于区分提供者是否为特定医疗设施；⑤医疗保健管理系统（HCPCS）标识，即服务是否包含在药品平均价格（ASP）文件中。

本文涉及的数值自变量见表 1。

表 1　　数值自变量及含义

变量名	变量含义
BUcnt	接受该医疗服务的日均保险受益人数
LScnt	医疗服务提供者的服务种类数
BDScnt	接受不同医疗服务的日均保险受益人数
AMAamt	医疗服务的平均保险允许金额
ASCamt	医疗服务的平均费用

（二）数据处理与可视化

上文 10 个感兴趣变量中后五个为数值变量，其余为分类变量。数值变量进行 $\log_{10}$ 变换和归一化处理后值域为 $[-1,1]$，清洗后数据集包含近 985 万条数据，数据文件大小超过 2GB。对完整数据建立复杂模型显然是低效的，故本文基于分治一步估计算法建立医疗保险平均支付金额的线性预测模型。为提高计算效率，结合渐近正态成立的条件 $s = O(\sqrt{N})$ 和研究的实际意义，根据提供者所在州将数据集划分成 51 个子数据集，其中数据量最大加利福尼亚州为 74.59万条；数据量最小的是阿拉斯加州 1.7 万条。

（三）自变量的确定

随机抽取完整数据集中 20% 的数据计算相关系数，根据相关性将感兴趣数值变量分为两类，第一类为医疗保险服务数等 3 个包含服务提供者信息的变量，第二类为医疗保险允许支付的平均金额等 3 个包含相应医疗保险信息的变量，两类之内高度相关，两类之间无明显相关性，表明医疗服务提供者特征对医疗保险平均支付金额影响微弱，建立预测模型时应更关注相应医疗服务的特性。

表 2　　数值变量相关系数

变量名	LScnt	BUcnt	BDScnt	AMAamt	ASCamt	AMPamt
LScnt	1.00	0.57	0.61	-0.01	-0.01	-0.01
BUcnt	0.57	1.00	0.97	-0.01	-0.01	-0.01
BDScnt	0.61	0.97	1.00	-0.01	-0.01	-0.01
AMAamt	-0.01	-0.01	-0.01	1.00	0.75	0.99
ASCamt	-0.01	-0.01	-0.01	0.75	1.00	0.74
AMPamt	-0.01	-0.01	-0.01	0.99	0.74	1.00

（四）分类变量的异质性

对于因变量标准化医疗保险平均支付金额，假定对数变换和归一化处理后的数值变量的影响是同质的，本文还感兴趣分类变量作用效果在各州的差异，因此假定分类变量存在异质性。为检验上述假设，分别建立局部数据集的线性预测模型，考察系数估计差异的显著性。

绘制分类变量在各州系数估计值的箱线图，由图 1 知，自变量提供者实体属性和 HCPCS 药物标识的各州系数估计值差异较大，异质性假设合理；提供者类型的 87 个虚拟变量对因变量的作用在 51 个子集内存在明显差异，可以认为异质性存在，不同虚拟变量的系数估计值分布特征也存在差异，分析时应更关注几个异质性明显且系数估计值显著不为 0 的类别，如：流感中心、公共免疫中心、公共卫生福利局等。

服务提供者医疗保险计划参与情况和服务机构类型系数估计值差异较小且接近 0，对比各州医疗保险计划参与情况系数估计的 P 值并与给定显著性水平 $\alpha=0.05$ 比较，发现该变量系数估计的显著性在不同子集中存在差异，从而体现出异质性，见图 2。

（五）数据分析

考察局部数据集上标准化医疗保险平均支付金额线性预测模型的拟合优度，汇总各州调整后 R^2 值，见表 3。结果表明各局部模型均具有良好的拟合效果。

为更好把握同质变量在美国全国范围内的作用，对各州系数估计结果进行聚合。由于数据集样本量差异较大，以局部加权估计作为初始估计量，应用分治一步加权估计算法，得到保险允许支付金额和医疗服务平均费用的系数估计值。绘制拟合直线，比较各州局部系数估计、加权平均估计和分治一步估计结果，见第 10 页图 3。

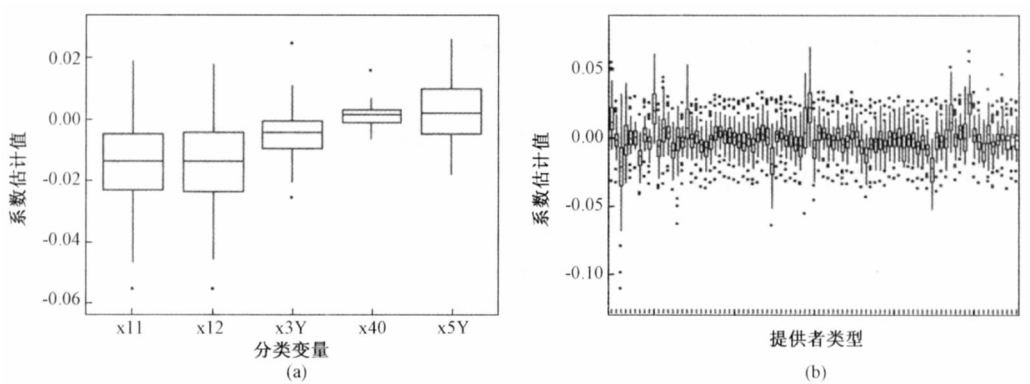

图 1　各州分类变量系数估计箱线图

注：图（a）横坐标分类变量 x11、x12、x3Y、x4O、x5Y 分别表示“提供者性别为男”“提供者性别为女”“提供者参与医疗保险计划”“服务机构为非设施”和“HCPCS 药物在 ASP 文件内”；纵坐标为各变量的系数估计；图（b）横坐标为“提供者类型”分类变量，纵坐标为系数估计。

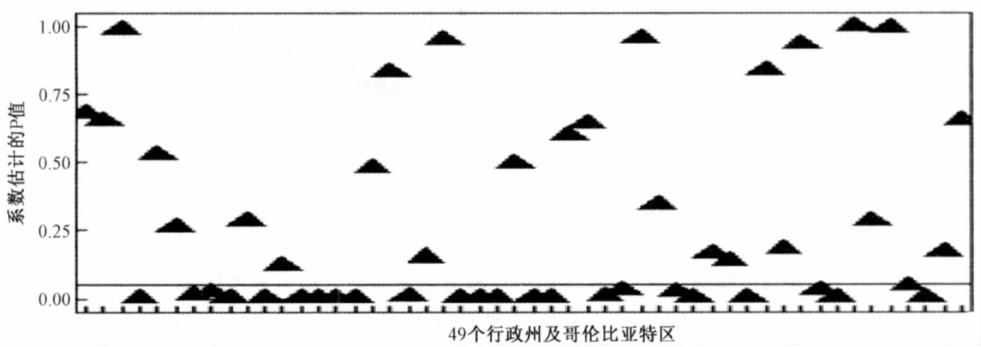

图 2　提供者参与医疗保险计划系数估计的 P 值

注:由于马萨诸塞州数据集内所有医疗服务提供者均参与了医疗保险计划,不构成分类变量,因此横坐标为除马萨诸塞州外 50 个州和特区,纵坐标为系数估计 P 值。

表 3　　美国各州标准化医疗保险平均支付金额线性模型拟合优度

州名	调整后 R^2	州名	调整后 R^2	州名	调整后 R^2
AK	0.9781	KY	0.9899	NY	0.9860
AL	0.9923	LA	0.9910	OH	0.9887
AR	0.9907	MA	0.9872	OK	0.9902
AZ	0.9902	MD	0.9984	OR	0.9876
CA	0.9883	ME	0.9870	PA	0 9880
CO	0.9872	MI	0.9892	RI	0.9901
CT	0.9878	MN	0.9890	SC	0.9907
DC	0.9887	MO	0.9891	SD	0.9860
DE	0.9901	MS	0.9921	TN	0.9905
FI	0.9904	MT	0.9825	TX	0.9906
GA	0.9908	NC	0.9901	UT	0.9869
HI	0.9866	ND	0.9840	VA	0.9880
IA	0.9893	NE	0.9880	VT	0.9836
ID	0.9873	NH	0.9854	WA	0.9872
II	0.9878	NJ	0.9893	WI	0.9866
IN	0.9892	NM	0.9860	WV	0.9883
KS	0.9883	NV	0.9902	WY	0.9923

首先,局部估计拟合直线的斜率几乎完全一致,表明相应数值变量对标准化平均支付金额的作用在各州之间没有明显差别,同质性假定合理;其次,加权系数估计的斜率明显偏离各州的共同斜率,难以很好地刻画同质变量的共性;而一步估计与局部估计拟合曲线的斜率几乎一致,说明了分治一步估计方法具有更快的收敛速率。

五、结果讨论

本文考虑了医疗服务提供者类型等分类变量对医疗保险平均支付金额影响的异质性,运用分治一步算法建立标准化美国医疗保险平均支付金额预测模型。借助差异对比图可以讨论分类变量在各州的不同影响。

图 4 反映了相对于提供者为实体,提供者为男

性和女性对标准化后医疗保险平均支付金额的影响。比较两变量系数估计可知,具体性别对因变量影响的差异不大,绝大多数州的系数估计值为负,说明相对于实体,医疗服务提供者为个人时,医疗保险的平均支付金额更少。这可能由于个人医疗服务提供者中包含了执业护士或护理人员,拉低了医疗保险的平均支付金额。

图5(a)体现了医疗服务提供者医疗保险计划参与情况对平均支付金额的影响。该变量系数的1标准差区间在多个州的结果包含0且宽度较大,表明在多个州内"提供者医疗保险参与情况"对保险平均支付金额的影响不显著;该虚拟变量在绝大多数州的系数估计为负数,表明相对于不完全参与医疗保险计划,完全参与的平均支付金额更少。考虑到在美国当前"管理式医疗"体系下,不完全参与医疗保险计划的医疗服务提供者的保险允许金额比完全参与的提供者低5%左右,但可以向保险受益人收取20%的共同保险费用之外的额外费用,即服务的平均费用提高,该结果也说明医疗保险计划起到了压缩平均支付金额的作用。

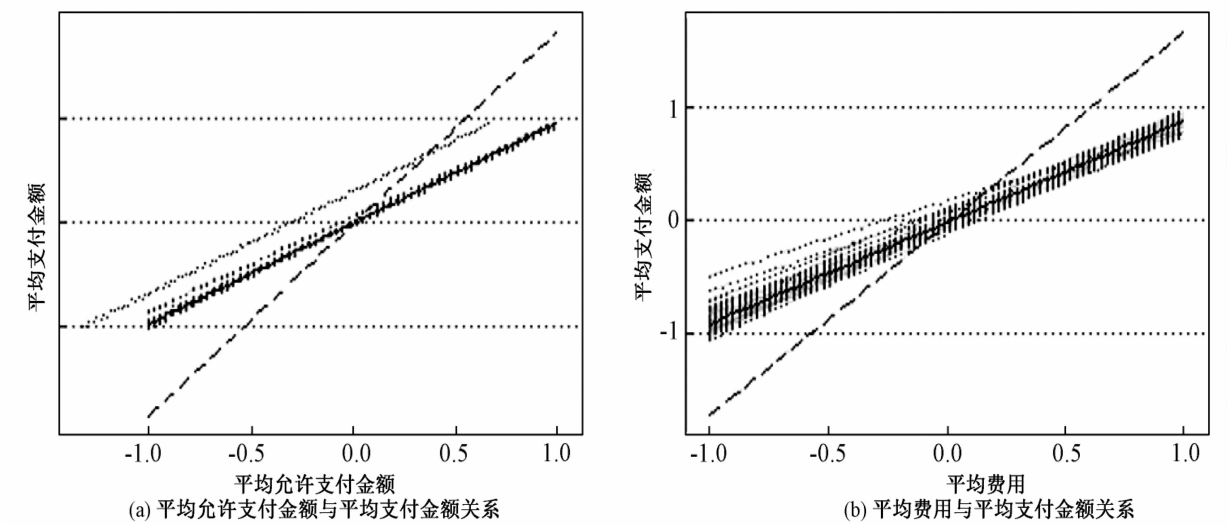

图3 局部估计、加权平均估计和分治一步估计拟合曲线比较

注:虚线表示51个行政州及特区数据的局部估计拟合结果,点线表示局部估计加权平均拟合结果,直线表示分治一步估计拟合结果。

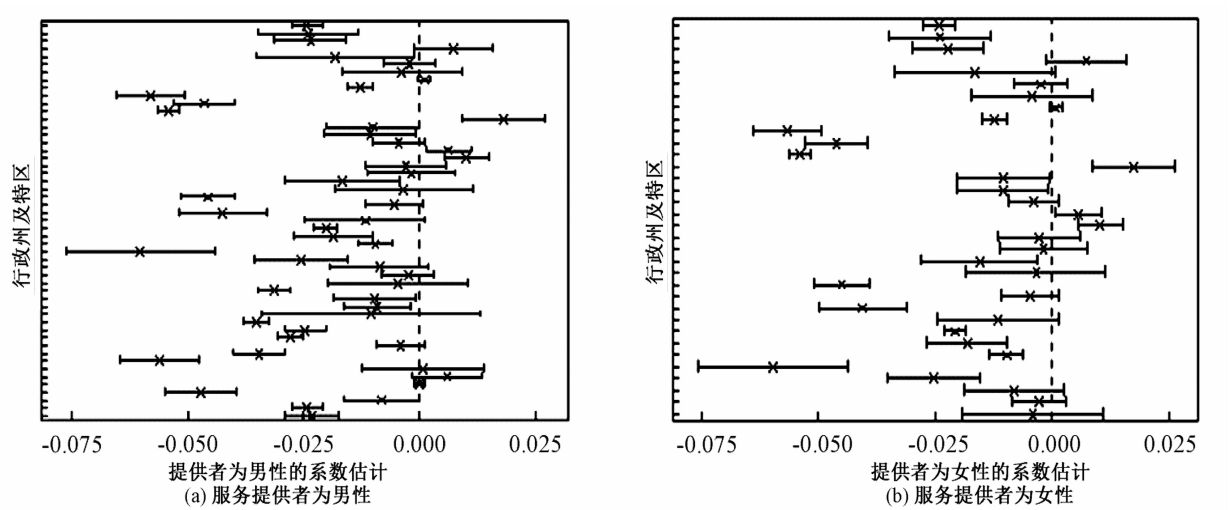

图4 医疗服务提供者属性影响差异对比图

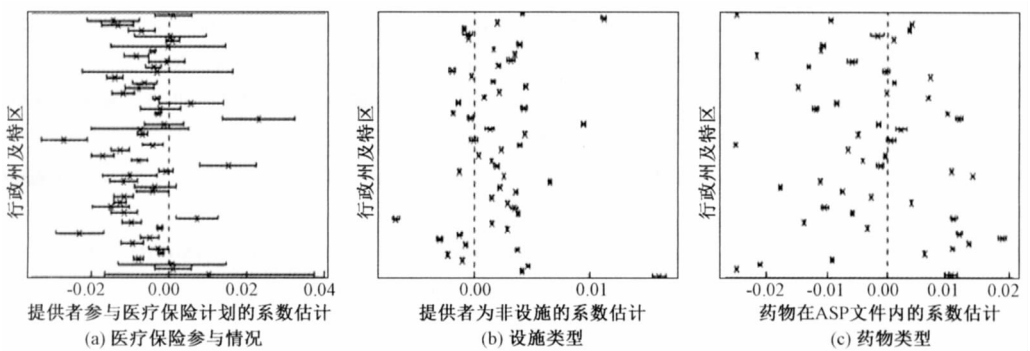

图 5　其他分类变量影响差异对比图

图 5(b)为服务设施类型对平均医疗保险支付金额影响在各州之间的差异对比图。该变量系数估计多为 0 附近且为较小正数,表明服务设施类型在各州的作用存在异质性,但本身作用较小导致在图像上差异不大;系数为正说明服务地点为非设施机构时,医疗保险的平均支付金额更高。这是因为医疗服务提供者为大型医疗设施时,受益人医疗服务和医疗设施使用费用通常按具体项目多次支付,因此在服务记录中单次支付的金额较少。

图 5(c)反映了 HCPCS 药物类型对因变量的影响。该变量在各州的系数估计值同样呈现出标准差小的特征;系数估计值在各州之间差异较大,体现出其作用的异质性。在绝大多数州变量的系数估计值为负,说明医疗服务和药品的 HCPCS 标识包含在 ASP 文件内时,医疗保险的平均支付金额更少,该结果引导医疗服务提供者优化服务方案以控制成本。

最后考虑提供者类型对平均支付金额的影响。在提供者类型的 87 个虚拟变量中,系数估计值显著为正且绝对值较大的变量对应的类别为:流感中心、公众免疫中心、公共卫生福利局、注册营养师或营养专业人士;系数估计显著为负且绝对值较大的变量对应的类别为:救护车服务提供者、听觉矫正专家、独立诊断测试设施、便携式 X 射线检查提供者。对比上述结果可以发现,前者主要提供综合性、长期性的医疗服务,后者对应的则是一次性、专业性的医疗服务,与其他项目关联性强,如救护车服务通常与后续的检查、诊治服务相关。

本文主要讨论了基于分治策略的海量异质性数据统计推断问题,提出了分治一步加权估计方法,证明了其渐近性质并考虑异质系数的假设检验问题。基于美国医疗保险使用和支付数据建立了全美各州标准化医疗保险平均支付金额的分治线性预测模型,结果说明分治一步估计具有更快的均方误差收敛速度。实证表明美国医疗保险体系的确存在覆盖范围有限,制度流程复杂的弱点,但对异质性分类变量的作用进行分析发现,在当前体系下,“管理式医疗”及其配套规则确实起到了控制医疗服务费用的作用,我国医疗保险制度改革中可借鉴吸取其经验教训,完善数据收集和管理制度,在监管中合理发挥大数据的作用。

参考文献:

[1]陈心洁,林鹏,邹国华. 线性混合效应模型的 FIC 选择准则[J]. 统计研究,2015,32(3):100 – 103.

[2]梅波,徐礼文. 基于极大似然的异方差多正态总体均值的参数 Bootstrap 检验[J]. 数理统计与管理,2016,35(4):630 – 640.

[3]苏鹏,田茂再. 基于最小化复合分位损失函数的尺度参数估计和异质性检验[J]. 系统科学与数学,2018,38(9):1055 – 1066.

[4]张新雨,邹国华. 模型平均方法及其在预测中的应用[J]. 统计研究,2011,28(6):97 – 102.

[5]Battey H, Fan J, Liu H, et al. Distributed Testing and Estimation under Sparse High Dimensional Models[J]. The Annals of Statistics, 2018, 46(3):1352 – 1382.

[6]Boyd S, Parikh N, Chu E, et al. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers[J]. Foundations & Trends in Machine Learning, 2010, 3(1):1 - 122.

[7]Chen X, Liu W, Zhang Y. Quantile Regression under Memory Constraint[J]. The Annals of Statistics, 2019, 47(6):3244 - 3273.

[8]Huang C, Huo X. A Distributed One - Step Estimator[J]. Mathematical Programming, 2019, 174:41 - 76.

[9]Lian H, et al. Projected Spline Estimation of the Nonparametric Function in High - Dimensional Partially Linear Models for Massive Data[J]. The Annals of Statistics, 2019, 47(5):2922 - 2949.

[10]Lin L, Li F. A Global Bias - Correction DC Method for Biased Estimation under Memory Constraint[R]. arXiv preprint arXiv:1904.07477, 2019.

[11]Lu J, Cheng G, Liu H. Nonparametric Heterogeneity Testing for Massive Data[R]. arXiv preprint arXiv:1601.06212, 2016.

[12]Shamir O, Srebro N, Zhang T. Communication Efficient Distributed Optimization Using an Approximate Newton - type Method[C]. International Conference on Machine Learning, 2014:1000 - 1008.

[13]Shi C, Lu W, Song R. A Massive Data Framework for M - Estimators with Cubic - Rate[J]. Journal of the American Statistical Association, 2018, 113(524):1698 - 1709.

[14]Wang B, Fang Y, Lian H, et al. Additive Partially Linear Models for Massive Heterogeneous Data[J]. Electronic Journal of Statistics, 2017, 13(1):391 - 431.

[15]Zhang Y, Duchi J C, Wainwright M. Communication - Efficient Algorithms for Statistical Optimization[J]. Journal of Machine Learning Research, 2013, 14(1):3321 - 3363.

[16]Zhao T, Cheng G, Liu H. A Partially Linear Framework for Massive Heterogeneous Data[J]. The Annals of Statistics, 2016, 44(4):1400 - 1437.

Distributed Estimation for Heterogeneous Big Data

Guo Jingxuan Xu Huichao Zhu Wanqing Tian Maozai

Abstract: With the rapid development of IoT technology, big data brings great challenge to network bandwidth and computer storage capacity, which makes traditional centralized data processing difficult to achieve. Distributed computing came into being in this background. The idea of distributed computing, known in statistics as divide - and - conquer(DC), is attracting more and more attention from statisticians. Zhang et al.(2013) demonstrated the simple average of local empirical risk minimization estimation has mean square error rate $O(N^{-1})$ when the number of data sets $s = O(\sqrt{N})$. On this basis, Huang and Huo(2019) proposed a distributed one - step estimator of M - estimation with Newton - Raphson iteration. However, the above methods do not consider the effect of heterogeneity in big data on estimation results. In this paper, a distributed one - step weight estimation for heterogeneous big data is proposed in the framework of linear model and its asymptotic properties are proved and used to test heterogeneity in big data. In addition, the proposed method is applied to the actual data analysis of medical insurance in the United States. The results show that compared with the simple average estimation, the method presented in this paper can better fit the linear trend of data and significantly improve the computational efficiency.

Key words: divide - and - conquer; one - step estimator; big data; heterogeneity; medical insurance