ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 34,No. 5,Sep. 2019,pp. 883－892
DOI:10. 16337/j. 1004-9037. 2019. 05. 013

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于层次结构数据的多元线性回归问题分析

赵　芸[1]　唐旭清[1,2]

(1. 江南大学理学院，无锡，214122;2. 无锡市生物计算工程技术研究中心，无锡，214122)

摘　要：针对传统多元线性回归分析(Multiple linear regression，MLR)在处理大数据时，特别是具有层次结构的数据，提出了基于层次结构数据的偏回归系数计算模型。该模型通过计算下层中每个部分的偏回归系数及上、下层之间的层次结构矩阵，来计算上层的总体偏回归系数。从理论研究和实际数据试验验证了在计算回归系数时新模型与传统MLR模型具有等效性。同时，新模型能有效解决隐私数据的保护问题，实现计算的并行处理，提高了大数据处理能力。

关键词：多元统计;回归分析;层次结构数据;隐私保护

中图分类号：O212　　**文献标志码：**A

Multiple Linear Regression Problem Based on Hierarchical Structure Data

Zhao Yun[1], Tang Xuqing[1,2]

(1. School of Science, Jiangnan University, Wuxi, 214122, China; 2. Wuxi Engineering Research Center for Biocomputing, Wuxi, 214122, China)

Abstract: Multiple linear regression (MLR) is widely used in statistical analysis. Based on common tools of the multiple linear regression in big data research, especially in the research of hierarchical structure data, a partial regression coefficient model is proposed here. The total partial regression coefficient is calculated by using each partial regression coefficient at the lower part and the hierarchical matrix between the lower and upper parts. It is validated that the new model is equivalent to the common models of multiple linear regression by the theoretical research and the real data. The new method can effectively solve the problem of privacy data in privacy protection research. Moreover, the new model can realize the parallel computation, which improves the capability of big data processing.

Key words: multivariate statistics; regression analysis; hierarchical structure data; privacy protection

引　　言

线性回归(Linear regression，LR)分析或多元线性回归(Multiple linear regression，MLR)分析[1-3]主要用于研究变量间的相关关系及基于数据变量间客观规律的获取。作为一种常用的统计分析方法，MLR在实际问题研究中得到了广泛应用[4-7]，同时理论也得到不断丰富和发展[3-8]。

近年来，随着计算机科学和网络技术的飞速发展，大批量数据不断涌现，大数据已经成为许多部门与行业一个重要的特点[9-11]。受实际需求影响，在大数据存储、计算过程中数据量庞大，一般多采用拓

基金项目：国家自然科学基金(11371174，11271163)资助项目;国际科技合作研究(2011DFR70500)资助项目。
收稿日期：2017-09-12;**修订日期：**2017-12-13

扑结构形式进行存储,其中较为常见的就是层次结构[12-13]。层次结构作为一种常用的数据结构,具有典型的树状特点,有利于存储数据的管理与检索。如银行、保险、医疗等行业的数据按行政区划就具有层次结构的特点,并且这些行业需要利用大数据处理技术进行不同地区间或者不同行业间的数据整合与分析。因此,基于层次结构的数据处理与计算技术研究就显得尤其重要和紧迫[13-14]。

随着大数据研究的不断深入,基于大数据的MLR模型被广泛应用于数据处理中。王慧文等[15]提出了MLR模型的增量算法,该算法可在已知全部数据信息的前提下,节约数据读取时间,减小了数据存储传输的压力。此外对于不同的回归分析模型,如Logistic回归也渐渐被引入大数据处理,并产生了相应的算法,Jiang等[16]提出了基于网络分布式数据的Logistic回归分析算法,用于数据间的规律获取。这些基于大数据处理与计算方法的探索与研究有利于提高计算的效率,同时对于具有层次结构的数据进行处理与计算时,除考虑现有问题外,更需要解决各层之间的联系以及数据综合的问题,如各分层部分的MLR系数与总的MLR系数的数量计算关系。除此之外,在一些特殊行业中,例如金融服务、医疗卫生等领域还面临着数据安全和隐私保护的问题,并已经成为大数据研究的重要问题之一。冯登国等[17]从宏观方面提出了大数据安全与隐私保护的一些构想。罗永龙等[18]提出了一种基于安全协议的隐私保护方法,并应用MLR分析方法进行研究。美国加州大学圣地亚哥分校的Jiang教授团队就分布式数据提出了隐私保护协议的支持向量机算法[19-20]。

在以上研究的基础上,本文提出了层次结构数据的MLR分析方法的研究,其主要目的是通过下层数据的部分偏回归系数以及层次结构矩阵来求解上层模型的偏回归系数,以此来实现由部分偏回归系数来构建全体MLR模型的目标。针对下层每个部分的偏回归系数,数据用户只需要提供原数据总和、平方和以及交叉项乘积和即可求解该部分的MLR模型的偏回归系数。与直接利用原始数据求解偏回归系数的相比,通过原数据总和、平均值以及交叉项乘积和的输入进行偏回归系数的求解,既可以保证原始数据的私密性,又可达到与原始数据直接输入相同的结果。同时模型可实现整个计算的并行处理,提高大数据处理能力。

1 基于层次结构数据的偏回归系数计算方法

1.1 带加密数据库的层次结构数据

在大数据分析处理中,为方便数据的存储、读取、计算等操作,大部分数据都按照一定拓扑结构进行存储,如链式结构、网状结构、环形结构等,其中较为常用的一种数据管理结构为层次结构。

通过层次结构所组成的数据即为层次结构数据[21],层次结构数据具体关系见图1。在层次结构数据中,所有数据点组成一个层次化的垂直树形网络,每一上层数据集拥有下层分支的全部数据成员。在实际操作过程中,对一个共含有P层的层次结构数据集合,第P层的各数据集将全部数据传输到该节点对应的上层数据节点,然后对第$P-1$层的各数据集汇总,并传输到其对应的第$P-2$层数据节点上,每次往上一层汇总时,会对汇总层进行置空,以此类推,直到传输汇总到第1层数据节点。

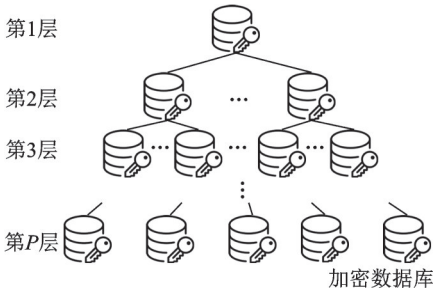


图1 层次结构拓扑图
Fig.1 Hierarchy topology

通过层次结构化的垂直树形网络,数据被逐层传递汇总,在实际的计算分析中数据既可以在当前数据层进行处理,也可以在上层进行汇总处理。这样既可以保持统计规律不改变,又实现了并行处理,增加了数据的灵活性和可用性。基于此特点,层次结构数据在银行、金融、医疗卫生[22]等行业领域

有着广泛的适用性。

同时在银行、金融、医疗卫生等行业领域中,数据集中往往包含着用户的隐私信息,因此多采用加密数据库进行存储、传输。在加密数据库中,每个数据库仅保留少量外部接口或对数据进行加密处理,两者加密方法都对基于全体数据的回归分析模型在构建上造成一定困难。为了对基于隐私数据的层次结构数据集进行回归分析,本文在传统回归分析的基础上,提出基于少量接口数据的回归数据计算方法,算法如下:

步骤1　开始;

步骤2　参数初始化p、P,令$p=P$;

步骤3　由第p层接口数据求解部分偏回归系数B_p、层次结构矩阵$\boldsymbol{Q}_p$,令$p=p-1$;

步骤4　由部分偏回归系数B_p、层次结构矩阵$\boldsymbol{Q}_p$,求解总体偏回归系数B;

步骤5　判断p值,如果$p>1$转步骤2,如果$p=1$转步骤6;

步骤6　结束。

在该算法中,参数p为计数器,计算当前所在的层数,参数P是层次结构数据的总层数。算法中步骤2负责计算包含少量接口数据的下层部分偏回归系数,在充分保护数据隐私的前提下构建结构下层数据中小部分数据的MLR模型。步骤3负责利用下层部分偏回归系数以及数据传递时的层次结构矩阵计算上层总体偏回归系数。在步骤2,3的计算过程中,所有偏回归系数以及层次结构矩阵的计算仅需少量接口数据,因此本文算法能在构建层次结构数据MLR模型的同时,充分保障数据的私密性。步骤3,4的具体计算方法如下。

1.2　层次数据的回归建模

考虑一组已知的层次结构数据,采用MLR分析对其结构内数据进行建模计算,由层次结构的特点,本文考虑对其中任意上下两层数据子集进行分析。该数据子集中上层有一个部分,下层由K个部分组成,数据上下层之间满足层次结构,且下层之间数据相互独立。在此数据集的基础上本文考虑构建上层总体偏回归系数与下层部分偏回归系数之间的关系模型。

1.2.1　部分偏回归系数计算

以下将具体阐述下层部分偏回归系数的求解方法。为达到保护隐私的目的,本文方法只需少量接口数据便可进行下层每个部分偏回归系数的求解,其中接口数据包括原数据总和、平均值及交叉项乘积和。

在传统MLR分析中,利用最小二乘求解方法[23]求解偏回归系数仅需计算

$$\begin{cases} L_{11}b_1+L_{12}b_2+\cdots+L_{1N}b_N=L_{10} \\ L_{21}b_1+L_{22}b_2+\cdots+L_{2N}b_N=L_{20} \\ \qquad\qquad\vdots \\ L_{N1}b_1+L_{N2}b_2+\cdots+L_{NN}b_N=L_{N0} \end{cases} \tag{1}$$

式中N表示回归模型中自变量X的维数。

对式(1)中回归系数方程组系数矩阵$\boldsymbol{L}$与$\boldsymbol{L}^0=\left(L_{10},L_{20},\cdots,L_{N0}\right)^{\mathrm{T}}$的计算方法通常如下

$$\begin{cases} L_{ij}=\sum\limits_n(x_{ni}-\bar{x}_i)(x_{nj}-\bar{x}_j) \\ L_{i0}=\sum\limits_n(x_{ni}-\bar{x}_i)(y_n-\bar{y}) \end{cases} \qquad i,j=1,2,\cdots,N \tag{2}$$

式中:N表示回归模型中自变量$\boldsymbol{X}$的维数;n表示每一维自变量的样本数。

在式(2)算法中需要已知全体自变量$\boldsymbol{X}$和应变量$\boldsymbol{Y}$的原始数值才可以进行计算求解。但在一些特定场合中,原始数据是严格保密的,因此本文考虑通过原数据总和、平均值、交叉项乘积这类不涉及隐私信息、可用于传输的接口数据来构造部分偏回归系数方程组系数矩阵$\boldsymbol{L}$和常数向量$\boldsymbol{L}^0$。

本文考虑对式(2)中的L_{ij}进行展开计算,以此来设计新的方程组系数构造方法。展开后结果如下

$$L_{ij}=T(x_i,x_j)-\bar{x}_iT(x_j)=T(x_i,x_j)-\bar{x}_jT(x_i) \tag{3}$$

其次本文考虑对L_{j0}进行展开计算,可得到

$$L_{i0}=T(x_i,y)-\bar{y}T(x_i) \tag{4}$$

式中:第i维自变量数据的平均值为$\bar{x}_i$;第i维自变量数据的总和为$T(x_i)$;与第j维自变量数据的交叉乘积和为$T(x_i,x_j)$;应变量数据的平均值为$\bar{y}$;因变量与第i维自变量数据的交叉乘积和为$T(x_i,y)$。

这样即可得到下层部分偏回归系数的两部分系数,非常数项偏回归系数

$$\boldsymbol{b}_i=\boldsymbol{A}\boldsymbol{L}^0 \quad i=1,2,\cdots,N \tag{5}$$

以及常数项偏回归系数

$$\boldsymbol{b}_0=\bar{\boldsymbol{y}}-\sum_{i=1}^{N}b_i\bar{\boldsymbol{x}}_i \tag{6}$$

式中N为自变量维数;在式(5)中的矩阵$\boldsymbol{A}$是由原数据总和、平均值、交叉项乘积所构造的系数逆矩阵,具体表达式为$\boldsymbol{A}=\boldsymbol{L}^{-1}$。$\boldsymbol{L}^0$的具体表达式为$\boldsymbol{L}^0=\left(L_{10},L_{20},\cdots,L_{N0}\right)^{\mathrm{T}}$;在(6)式中$\bar{\boldsymbol{y}}$为因变量的平均值,$\bar{\boldsymbol{x}}_i$为第$i$维自变量的平均值。

通过上述求解推导,本文旨在对于原有回归分析的求解方法做进一步展开合并计算,并通过原数据总和、平均值、交叉项乘积来构造式(1)方程组中的系数$\boldsymbol{L}$,以此来求解部分偏回归系数$\boldsymbol{B}=[b_1,b_2,b_3,\cdots,b_N]$。同时在方程组求解过程中又引入系数逆矩阵$\boldsymbol{A}$来替代原有的$\boldsymbol{L}$,进一步化简的偏回归系数求解方法。

1.2.2 总体偏回归系数计算

本节将构建上层总体偏回归系数与下层部分偏回归系数之间的关系模型。

考虑MLR分析中最小二乘的矩阵求解方法

$$\boldsymbol{B}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \tag{7}$$

在本文模型对应的层次结构数据中,式(7)中的$\boldsymbol{X}$、$\boldsymbol{Y}$包含了K个数据部分,第k部分的数据为$\boldsymbol{X}^{(k)}$和$\boldsymbol{Y}^{(k)}$($k=1,2,\cdots,K$),由模型的线性可加性可知,式(7)中的$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$、$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$可表示为

$$\begin{cases}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}=\sum\limits_{k=1}^{K}\boldsymbol{X}^{(k)\mathrm{T}}\boldsymbol{X}^{(k)}\\ \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}=\sum\limits_{k=1}^{K}\boldsymbol{X}^{(k)\mathrm{T}}\boldsymbol{Y}^{(k)}\end{cases} \tag{8}$$

由最小二乘法的矩阵表示形式可知,式(8)中的$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$可表示成

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}=\left(\boldsymbol{X}^{(k)\mathrm{T}}\boldsymbol{X}^{(k)}\right)\boldsymbol{B}_k \tag{9}$$

将式(8,9)代入式(7),可得第k部分结构数据的偏回归系数$\boldsymbol{B}_k$与总体偏回归系数$\boldsymbol{B}$之间的关系为

$$\boldsymbol{B}=\left(\sum_{k=1}^{K}\boldsymbol{X}^{(k)\mathrm{T}}\boldsymbol{X}^{(k)}\right)^{-1}\left(\left(\boldsymbol{X}^{(k)\mathrm{T}}\boldsymbol{X}^{(k)}\right)\boldsymbol{B}_k\right) \tag{10}$$

将$\boldsymbol{X}^{(k)\mathrm{T}}\boldsymbol{X}^{(k)}$表示为层次结构矩阵$\boldsymbol{Q}_k$,进行展开计算后可得到

$$\boldsymbol{Q}_k=\begin{bmatrix} N & T(\boldsymbol{X}_1^{(k)}) & T(\boldsymbol{X}_2^{(k)}) & \cdots & T(\boldsymbol{X}_N^{(k)}) \\ T(\boldsymbol{X}_1^{(k)}) & T(\boldsymbol{X}_1^{(k)},\boldsymbol{X}_1^{(k)}) & T(\boldsymbol{X}_1^{(k)},\boldsymbol{X}_2^{(k)}) & \cdots & T(\boldsymbol{X}_1^{(k)},\boldsymbol{X}_N^{(k)}) \\ T(\boldsymbol{X}_2^{(k)}) & T(\boldsymbol{X}_1^{(k)},\boldsymbol{X}_2^{(k)}) & T(\boldsymbol{X}_2^{(k)},\boldsymbol{X}_2^{(k)}) & \cdots & T(\boldsymbol{X}_2^{(k)},\boldsymbol{X}_N^{(k)}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ T(\boldsymbol{X}_N^{(k)}) & T(\boldsymbol{X}_1^{(k)},\boldsymbol{X}_N^{(k)}) & T(\boldsymbol{X}_2^{(k)},\boldsymbol{X}_N^{(k)}) & \cdots & T(\boldsymbol{X}_N^{(k)},\boldsymbol{X}_N^{(k)}) \end{bmatrix} \tag{11}$$

式中:$\boldsymbol{X}^{(k)}$为第k部分数据矩阵的扩展矩阵,即第1列数据全为1,第2列到最后一列为原始数据;第k部分数据的第i个分量的总和为$T(\boldsymbol{X}_i^{(k)})$,平方和为$T(\boldsymbol{X}_i^{(k)},\boldsymbol{X}_i^{(k)})$;与第$j$个分量的平方和为$T(\boldsymbol{X}_i^{(k)},\boldsymbol{X}_j^{(k)})$。

通过式(11)的计算方法,直接输入数据可得到$\boldsymbol{Q}_k$,结合计算$\boldsymbol{B}_k$可以得到总体偏回归系数

$$\boldsymbol{B}=\left(\sum_{k=1}^{M}\boldsymbol{Q}_k\right)^{-1}\left(\sum_{k=1}^{M}\boldsymbol{Q}_k\boldsymbol{B}_k\right) \tag{12}$$

式中:$\boldsymbol{B}_k$为下层第k部分数据的偏回归系数;$\boldsymbol{B}$为上层全体数据的总体偏回归系数。

基于式(12),可通过部分偏回归系数以及层次结构间的矩阵来计算任意p层与$p-1$层之间满足层次结构数据关系的偏回归系数。当层次结构数据由下往上按图1方式传输时,任意2层之间满足关系的数据就可构建上下层之间的偏回归系数模型,由此就可构建整个层次结构数据的偏回归系数关系模型。这种新的数据处理模式,对于具有层次结构的大数据处理具有重要意义。在不影响规律提取的前提下,一方面数据的分块处理能有效保护数据的隐私性;另一方面数据能分块处理可实现计算机的并行运算,提高大数据处理的能力。此外,通过理论推导可知本文的模型计算均为精确值。但在实际计算中,计算工具会导致截断误差的存在,不影响模型结果。

2　数据模型验证

在经济学研究中,多元性回归分析是一种常用的方法。本文参考韩琴等[24]在2017年提出的财政收入MLR模型,建立起2015年我国财政收入Y与人口数X_1、最终消费支出X_2、农业总产值X_3、工业总产值X_4、建筑业增加值X_5、灾害直接经济损失X_6之间的MLR方程,通过财政收入的MLR方程来验证本文所提方法模型的准确性。

同时为使数据呈现层次结构,本文将全国31个省市地区按照孙红玲等[25]提出的中国经济区的横向划分方法将全国31个省市地区划分为泛珠三角经济区、泛长三角经济区、大环渤海经济区,同时每个经济区分别包含12、10和9个省市地区,本文通过此经济区域划分来构建层次结构数据。具体结构如图2所示。

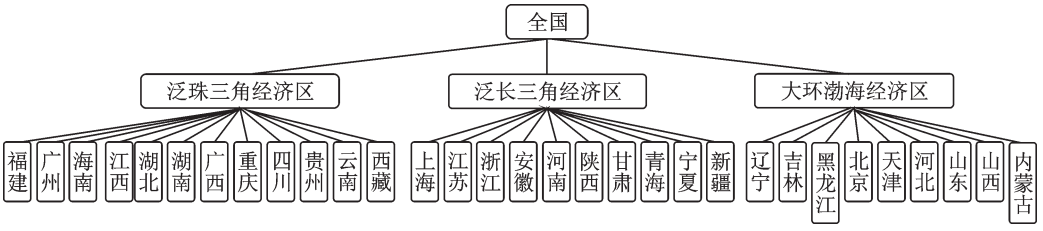


图2　基于横向划分的全国经济区域层次结构图

Fig.2　The Horizontal Structure of National Economy Based on Horizontal Division

根据图2结构,参照回归模型中所需的自变量和因变量,从2016年统计年鉴[26]中可知表1列出的数据。

通过表1数据及相应的层次结构,进行回归系数模型的验证。在验证过程中,将本文模型所求结果

表1 2015年全国31个省市地区统计数据

Tab. 1 Statistics of 31 provinces and cities in the country in 2015

经济区	地区	财政收入 Y / 亿元	人口数 X_1 / 亿人	最终消费支出 X_2 / 亿元	农业总产值 X_3 / 亿元	工业资产总计 X_4 / 亿元	建筑业增加值 X_5 / 亿元	灾害直接经济损失 X_6 / 亿元
泛珠三角经济区	福建	2 544.24	0.383 9	1 0328.9	3 717.9	17 240	2 270.876 2	189.1
	广东	9 366.78	1.084 9	3 7211.3	5 520	42 113	1 962.722 5	315.3
	海南	627.7	0.091 1	2 242.7	1 323.9	380	52.987 2	14.2
	江西	2 165.74	0.456 6	8 418.3	2 859.1	9 941	841.221	69.7
	湖北	3 005.53	0.585 2	13 799.7	5 728.6	16 413	1 960.367 9	82.2
	湖南	2 515.43	0.678 3	14 755.8	5 630.7	13 992	1 277.135 7	126.7
	广西	1 515.16	0.479 6	8 878.5	4 197.1	5 518	535.187 2	48.5
	重庆	2 154.83	0.301 7	7 503.2	1 738.1	6 608	1 501.246 9	22
	四川	3 355.44	0.820 4	15 775	6 377.8	13 525	1 373.791 6	132
	贵州	1 503.38	0.353	5 957.7	2 738.7	4 482	274.897 8	72.6
	云南	1 808.15	0.474 2	8 855.3	3 383.1	3 876	551.653 9	141.9
	西藏	137.13	0.324	820	149.5	104	23.736 6	107.2
泛长三角经济区	上海	5 519.5	0.241 5	14 854.5	302.6	8 994	834.169 4	3.5
	江苏	8 028.59	0.797 6	3 5041.4	7 030.8	48 488	5 996.522 3	84.9
	浙江	4 809.94	0.553 9	20 936.3	2 933.4	41 167	4 713.135 6	228.2
	安徽	2 454.3	0.614 4	10 970.5	4 390.8	19 077	1 206.511	118.9
	河南	3 016.05	0.948	18 722.6	7 641.3	22 892	1 618.362 9	44
	陕西	2 059.95	0.379 3	8 200	2 813.5	5 413	903.273 7	72.7
	甘肃	743.86	0.26	4 374.2	1 722.1	2 148	328.033 8	61.6
	青海	267.13	0.588	1 486	319.3	575	79.536 3	12
	宁夏	373.45	0.668	1 719.7	483	1 245	91.935 6	8.3
	新疆	1 330.85	0.236	5 639.8	2 804.4	2 707	532.662 5	155.8
大环渤海湾经济区	辽宁	2 127.39	0.438 2	13 019.5	4 686.7	12 304	10 653.841	65.1
	吉林	1 229.35	0.275 3	5 593.2	2 880.6	5 682	374.162 6	81.9
	黑龙江	1 165.88	0.381 2	8 986.7	5 044.9	4 162	247.110 2	39.5
	北京	4 723.86	0.217 1	14 503.6	368.2	3 548	1 237.418 5	1.3
	天津	2 667.11	0.154 7	7 155.7	467.4	5 525	663.367 3	0
	河北	2 649.18	0.742 5	13 197.8	5 978.9	15 295	754.963 5	107.5
	山东	5 529.33	0.984 7	26 144.4	9 549.6	41 485	2 078.520 8	80.7
	山西	1 642.35	0.366 4	7 134.7	1 522.6	3 845	461.340 3	103.3
	内蒙古	1 964.48	0.251 1	7 452.8	2 751.6	4 404	252.424 2	113.5

与Matlab自带工具箱求解结果进行比较,以此作对比验证。

对于3个经济区的数据,采用少量接口数据求解每部分的偏回归系数,再通过文中基于接口数据求得的部分偏回归系数以及层次结构矩阵,求解总体偏回归系数。

在求解部分偏回归系数时,本文假设表1数据集中的3个经济区的具体数值是未知,仅知道3个经济区数据总和、平方和以及交叉项乘积和,具体数值如表2—4所示。

表2　泛珠三角经济区接口数据表

Tab. 2　Interface data of the Pan-Pearl River Delta

变量	总和	平均值	交叉项乘积和					
			X_1	X_2	X_3	X_4	X_5	X_6
X_1	6.032 9	0.502 74	3.792 265 77	92 499.938 81	26 402.828 86	95 165.987 1	7 606.548 557	847.667 53
X_2	134 546.4	11 212.2	92 499.938 81	2 473 989 180	630 296 877.5	2 635 686 307	193 818 903.2	21 763 392.67
X_3	43 364.5	3 613.71	26 402.828 86	630 296 877.5	199 023 151.3	644 602 021.9	56 413 766.11	5 624 475.21
X_4	134 192	11 182.7	95 165.987 1	2 635 686 307	644 602 021.9	2 927 015 012	21 506 0657.4	23 443 377
X_5	12 625.824 5	1 052.15	7 606.548 557	193 818 903.2	56 413 766.11	215 060 657.4	20 001 661.85	1 771 716.285
X_6	1 321.4	110.117	847.667 53	21 763 392.67	5 624 475.21	23 443 377	1 771 716.285	220 200.82
Y	30 699.51	2 558.29	21 211.999 95	580 694 826.2	141 327 918.9	626 362 607.1	45 191 165.18	5 104 318.292

表3　泛长三角经济区接口数据表

Tab. 3　Interface data of the Pan-Yangtze River Delta

变量	总和	平均值	交叉项乘积和					
			X_1	X_2	X_3	X_4	X_5	X_6
X_1	5.286 7	0.528 67	3.336 617 07	75 223.371 36	19 934.451 51	101 491.248 8	10 532.161 47	402.685 94
X_2	121 945	12 194.5	75 223.371 36	2 481 118 697	551 237 723.9	3 404 500 223	37 6851 449.2	11 709 233.19
X_3	30 441.2	3 044.12	19 934.451 51	551 237 723.9	154 878 452.8	750 383 213.4	78 572 092.66	2 881 048.05
X_4	152 706	15 270.6	101 491.248 8	3 404 500 223	750 383 213.4	5 057 798 634	559 547 867.7	17 782 748.9
X_5	16 304.143 1	1 630.41	10 532.161 47	376 851 449.2	78 572 092.66	559 547 867.7	64 164 549.8	1 972 805.204
X_6	789.9	78.99	402.685 94	11 709 233.19	2 881 048.05	17 782 748.9	1 972 805.204	108 935.09
Y	28 603.62	2 860.36	16 463.286	576 108 276.5	117 124 549.9	769 777 029.2	86 129 103.45	2 632 328.085

表4　大环渤海经济区接口数据表

Tab. 4　Interface data of Circum-Bohai-Sea region

变量	总和	平均值	交叉项乘积和					
			X_1	X_2	X_3	X_4	X_5	X_6
X_1	3.811 2	0.423 47	2.202 427 78	54 955.710 59	20 013.731 77	64 888.879 5	8 076.683 609	292.046 53
X_2	103 188.4	11 465.4	54 955.710 59	1 507 270 247	491 099 372.5	1 667 084 989	235 193 362.9	6 791 005.75
X_3	33 250.5	3 694.5	20 013.731 77	491 099 372.5	192 899 782	604 503 233.3	78 781 391.69	2 623 753.17
X_4	96 250	10 694.4	64 888.879 5	1 667 084 989	604 503 233.3	2 233 231 204	242 954 921.1	7 324 452.1
X_5	16 723.148 4	1 858.13	8 076.683 609	235 193 362.9	78 781 391.69	242 954 921.1	120 843 421.3	1 060 780.267
X_6	592.8	65.866 7	292.046 53	6 791 005.75	2 623 753.17	7 324 452.1	1 060 780.267	54 129.44
Y	23 698.93	2 633.21	11 660.033 1	338 531 897.4	98 927 562.15	354 378 872.7	45 774 043.91	1 414 997.148

在表2—4中,总和与均值可以通过表1数据简单计算得出。而交叉项乘积和是需要进行计算的。通过表2—4中的数据,利用部分偏回归系数的求解方法可以将3个经济区每部分的偏回归系数计算出来。进而利用层次结构矩阵$\boldsymbol{Q}_k$构建的总体偏回归系数的求解方法去求解全国31个省市地区的总体偏

回归系数,结果如表5所示。

表5 部分偏回归系数、全体偏回归系数以及Matlab工具箱计算结果

Tab. 5 Partial regression coefficients, total partial regression coefficients, and Matlab toolbox calculations

	泛珠三角经济区偏回归系数	泛长三角经济区偏回归系数	大环渤海经济区偏回归系数	全体偏回归系数	Matlab工具箱结果
b_0	394.345 329 1	31.392 394 27	540.321 111 1	564.622 4	564.622 4
b_1	−403.504 134 7	−340.329 856 5	−847.046 364 4	−985.298 734 2	−985.298 734 2
b_2	0.265 595 748	0.422 868 039	0.304 947 64	0.319 282 237	0.319 282 237
b_3	−0.237 221 489	−0.421 952 338	−0.347 088 955	−0.244 025 894	−0.244 025 894
b_4	0.024 511 385	−0.015 032 01	0.033 679 247	−0.003 819 976	−0.003 819 976
b_5	0.118 778 312	−0.544 428 536	−0.077 057 037	−0.096 351 953	−0.096 351 953
b_6	−1.388 830 225	3.214 241 245	0.312 304 55	−0.904 322 944	−0.904 322 944

表5中的b_i表示每一维自变量的偏回归系数。本文模型求解的全体偏回归系数与Matlab工具箱结果相比,两者结果之间的计算误差数量级为10^{-11}到10^{-13}之间,属于Matlab工具本身导致的截断误差,不影响模型及方法本身,因此两者方法本身并无差距,由此可说明本文的总体偏回归系数模型有效可靠。

上述基于中国经济区的横向划分方法的31个省份财政收入的回归模型研究中,充分说明了本文提出的部分偏回归系数模型,以及基于层次结构矩阵的全体偏回归系数模型在实际应用中是可行、有效的。本文模型方法可在只提供原数据总和、平均值、交叉项乘积和等接口数据的前提下实现部分偏回归系数以及全体偏回归系数的求解,可适用于银行、医疗等领域在保护数据隐私前提下构建不同层次的回归分析模型。

3 结束语

本文针对大数据环境下海量的数据集以及数据处理的隐私保护问题,提出了基于层次结构矩阵来构建下层部分偏回归系数与上层总体偏回归系数之间关系的模型。理论推理表明模型可以利用原数据总和、平均值、交叉项乘积和这些带隐私保护功能的接口数据来求解部分偏回归系数。同时利用带隐私保护的接口数据求解层次结构矩阵,使层次结构矩阵也带有隐私保护功能,再通过部分偏回归系数以及层次结构矩阵求解总体偏回归系数,实现了全局模型的数据隐私保护。

同时以经济统计试验数据为例,验证了新模型的准确性。本文模型是对MLR模型及偏回归系数估计做出的有益的尝试,为大数据处理提供了更为快捷的方法,适用于不同行业的数据。同时,对于一些特殊行业的数据保密和隐私保护具有重要意义。

参考文献:

[1] Galton F. Regression towards mediocrity in hereditary stature [J]. Journal of the Anthropological Institute of Great Britain & Ireland, 1886, 15:246-263.

[2] Eberly L E. Multiple linear regression [J]. Methods in Molecular Biology, 2007, 404(2):165-187.

[3] 王惠文, 孟洁. 多元线性回归的预测建模方法[J]. 北京航空航天大学学报, 2007, 33(4):500-504.
Wang Huiwen, Meng Jie. Predictive modeling on multivariate linear regression [J]. Journal of Beijing University of Aeronautics and Astronautics, 2007, 33(4):500-504.

[4] Savescu R F, Laba M. Multivariate regression analysis applied to the calibration of equipment used in pig meat classification in

Romania [J]. Meat Science, 2016, 116:16-25.

[5] Hoeflinger J L, Hoeflinger D E, Miller M J. A dynamic regression analysis tool for quantitative assessment of bacterial growth written in Python [J]. Journal of Microbiological Methods, 2017, 132:83-85.

[6] Brix K V, Deforest D K, Tear L M, et al. Use of multiple linear regression models for setting water quality criteria for Copper: A complementary approach to the biotic ligand model [J]. Environmental Science & Technology, 2017, 51(9):5182-5192.

[7] Olaya-Abril A, Parras-Alcántara L, Lozano-García B, et al. Soil organic carbon distribution in mediterranean areas under a climate change scenario via multiple linear regression analysis [J]. Science of the Total Environment, 2017, 592:134.

[8] 张峰,陈华伟,李妍文. 基于多核最小二乘支持向量回归的TDOA-DOA映射方法[J]. 数据采集与处理,2017, 32(3):540-549.
Zhang Feng, Chen Huawei, Li Yanwen. TDOA-DOA mapping using multi-kernel least-squares support vector regression [J]. Journal of Data Acquisition and Processing, 2017, 32(3):540-549.

[9] Walker S J. Big data: A revolution that will transform how we live, work, and think [J]. American Journal of Epidemiology, 2014, 17(9):181-183.

[10] 吉根林, 赵斌. 时空轨迹大数据模式挖掘研究进展[J]. 数据采集与处理, 2015, 30(1):47-58.
Ji Genlin, Zhao Bin. Research progress in pattern mining for big spatio-temporal trajectories [J]. Journal of Data Acquisition & Processing, 2015, 30(1):47-58.

[11] 亓峰, 唐晓璇, 邢宁哲,等. 未来大数据环境下的配用电通信网虚拟网络架构及应用[J]. 数据采集与处理, 2015, 30(3): 511-518.
Qi Feng, Tang Xiaoxuan, Xing Ningzhe, et al. Virtual network architecture and application for smart distribution grid in future large data environment [J]. Journal of Data Acquisition & Processing, 2015, 30(3):511-518.

[12] 张铃, 张钹. 问题求解理论及应用——商空间粒度计算理论及应用[M]. 北京: 清华大学出版社, 2007.
Zhang Ling, Zhang Bo. Theory of problem solving and its applications—The theory and methods of quotient space of granular computing [M]. Beijing: Tsinghua University Press, 2007.

[13] 唐旭清, 朱平. 后基因组时代生物信息学的发展趋势[J]. 生物信息学, 2008, 6(3): 142-145.
Tang Xuqing, Zhu Ping. Developing trend of bioinformatics in post genome era[J]。 China Journal of Bioinformatics, 2008, 6 (3): 142-145.

[14] Tang Xuqing, Zhu Ping. Hierarchical clustering problems and analysis of fuzzy proximity relation on granular space [J]. IEEE Transactions on Fuzzy Systems, 2013, 21(5):814-824.

[15] 王惠文, 魏嫄, 黄乐乐. 多元线性回归模型的增量算法[J]. 北京航空航天大学学报, 2014, 40(11):1487-1491.
Wang Huiwen, Wei Yuan, Huang Lele. Incremental algorithm of multiple linear regression model [J]. Journal of Beijing University of Aeronautics and Astronautics, 2014, 40(11):1487-1491.

[16] Jiang X Q, Wu Y, Marsolo K, et al. Development of a web service for analysis in a distributed network [J]. eGEMs, 2014, 2 (1):1053.

[17] 冯登国, 张敏, 李昊. 大数据安全与隐私保护[J]. 计算机学报, 2014, 37(1):246-258.
Feng Dengguo, Zhang Min, Li Hao. Big data security and privacy protection [J]. Chinese Journal of Computers, 2014, 37(1): 246-258.

[18] 罗永龙, 徐致云, 黄刘生. 多元线形回归分析中的隐私保护问题[J]. 计算机工程与应用, 2005, 41(34):111-113.
Luo Yonglong, Xu Zhiyun, Huang Liusheng. Privacy protection in the multivariate linear regression problem [J]. Computer Engineering and Applications, 2005, 41(34):111-113.

[19] Yu H, Vaidya J, Jiang X. Privacy-preserving SVM classification on vertically partitioned data [C]// Advances in Knowledge Discovery and Data Mining. [S.l.]: PAKDD, 2006, 3918: 647-656.

[20] Vaidya J, Yu H, Jiang X. Privacy-preserving SVM classification [J]. Knowledge & Information Systems, 2008, 14(2):161-178.

[21] 夏纯中, 宋顺林. 基于商空间的层次式数据网格资源调度算法[J]. 通信学报, 2013, 34(6):146-155.
Xia Chunzhong, Song Shunlin. Hierarchical data grid resource allocation based on quotient space theory [J]. Journal on Communications, 2013, 34(6):146-155.

[22] 李晓松, 倪宗瓒. 对医学领域层次结构数据拟合线性回归模型时几个问题的探讨[J]. 四川大学学报(医学版), 1999, 30(1): 59-61.

Li Xiaosong, Ni Zongzan. On the problems of fitting linear regression models for hierachically structured data in medical research [J]. Journal of Sichuan University (Medical Science Edition), 19991, 30(1):59-61.

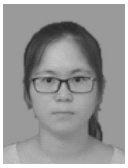
[23] 马立平. 回归分析[M]. 北京: 机械工业出版社, 2014.
Ma Liping. Regression analysis [M]. Beijing: China Machine Press, 2014

[24] 韩琴, 刘欢. 回归分析在我国财政收入中的应用[J]. 经贸实践,2017,(08):32-33.
Han Qin, Liu Huan. The application of regression analysis in China′s financial revenue [J]. Economic & Trade, 2017, (8): 32-33.

[25] 孙红玲,刘长庚. 论中国经济区的横向划分[J]. 中国工业经济, 2005, (10):29-36.
Sun Hongling, Liu Changgeng. The division in breadth of Chinese districts [J]. China Industrial Economics, 2005, (10):29-36.

[26] 中华人民共和国国家统计局. 中国统计年鉴—2016[M]. 北京:中国统计出版社, 2016.
National Bureau of the People′s Republic of China. China statistical yearbook—2016 [M]. Beijing: China Statistics Press, 2016.

作者简介:



赵芸(1991-),女,硕士研究生,研究方向:智能计算、生物信息学, E-mail: ljj22700@163.com。

唐旭清(1963-),男,教授,研究方向:智能计算、生物信息学、生态系统建模与仿真, E-mail: txq5139@jiangnan.edu.cn。

(编辑:张彤)