文章编号:1671-8860(2008)05-0504-04　　文献标志码:A

基于整体最小二乘法的线性回归建模和解法

鲁铁定[1,2]　陶本藻[1]　周世健[2,3]

(1　武汉大学测绘学院,武汉市珞喻路 129 号,430079)
(2　东华理工大学地球科学与测绘工程学院,抚州市学府路 56 号,344000)
(3　江西省科学院 ,南昌市上坊路,330029)

摘　要:对基于自变量和因变量误差的回归问题进行了进一步研究,证明了两种方法的实质并未解决同时考虑自变量和因变量的误差问题,其解算结果和不考虑自变量误差的解算结果完全相同。给出了能同时顾及自变量和因变量误差的新的回归模型,并推导了具体的解算方法。算例结果和基于矩阵分解的整体最小二乘法解算方法的结果相同,说明了本文方法的正确性。

关键词:线性回归;整体最小二乘;测量平差
中图法分类号:P207.2

近年来,围绕拟合方面的优选问题陆续有相关的研究成果[1,2],但这些方法不能很好地考虑自变量和因变量误差的影响。如果同时考虑回归模型中自变量和因变量的误差,其本质就是要在解算中考虑系数矩阵的误差,而常用的回归模型中一般认为系数矩阵是没有误差的[3-6]。测量数据处理中的平差模型、函数模型的系数通常是不考虑其误差的。如果需要顾及其误差,就可引用数学界已有的成果,结合测量模型实际,展开整体最小二乘平差(回归)(total least squares,TLS)的研究。本文基于 TLS 对一元线性回归模型进行了研究,与数学中的整体最小二乘方法不同,本文的推导沿用了测量平差中函数模型的建立和解算[3,4],使这一问题的解决更易于被求解和应用。

设模型的数据为$(x_i,y_i)(i=1,2,\cdots,n)$,在建模时,要解决的两个问题是:① 同时考虑 x_i 和 y_i 的误差;② 所建立的 y 关于 x、x 关于 y 的回归模型的解算结果应是一致的。这个问题看起来是一个简单的问题,其实不然。在测绘界,很早就对此类问题进行了研究,并提出了一些解法。文献[3]是同时顾及 x 和 y 误差的条件平差法;文献[4]是同时顾及 x 和 y 误差的间接平差法,特别是对于文献[4]中所提出的方法,在目前的 GIS 空间数据处理中已被多次引用[5,6]。基于上述两种方法所建立的模型看起来似乎已经考虑了系数的误差,但本文通过证明,这两种模型实际并未达到预定的目标,而且其平差结果与不考虑系数误差的普通最小二乘回归完全一致。本文通过对此两个模型存在问题的分析,建立了基于整体最小二乘法的线性回归新模型。理论和实践证明,该模型的平差结果解决了建模时的上述两个问题。与数学解法不同的是,本文将整体最小二乘解法纳入了测量平差方法的范畴,其成果为进一步研究测量函数模型的整体最小二乘法打下了研究基础。

1　顾及自变量误差的一元线性回归

1.1　一元线性回归问题

以 x 作 y 的一元线性回归模型为:

收稿日期:2008-03-16。
项目来源:国家自然科学基金资助项目(40574008);江西省自然科学基金资助项目(0711008,0650007);江西省教育厅科技基金资助项目(赣教财 2006[208]);武汉大学地球空间环境与大地测量教育部重点实验室开放研究基金资助项目(06-06,04-01-07);数字国土江西省重点实验室开放研究基金资助项目(DLLJ200506);地理空间信息工程国家测绘局重点实验室开放研究基金资助项目。

$$y_i = a + bx_i + \varepsilon_i, i = 1,2,\cdots,n \tag{1}$$

设 v_i 为 $-\varepsilon_i$ 的最小二乘估值,取回归参数 a、b 的近似值为 a_0、b_0,则式(1)可表示为:

$$v_i = \delta a + x_i\delta b - (-a_0 - b_0x_i + y_i), \quad i = 1,2,\cdots,n \tag{2}$$

其矩阵形式表示为:

$$\underset{n,1}{\boldsymbol{V}} = \underset{n,2}{\boldsymbol{A}}\underset{2,1}{\delta\boldsymbol{B}} - \underset{n,1}{\boldsymbol{W}} \tag{3}$$

式中,$\underset{n,2}{\boldsymbol{A}} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$,$\underset{2,1}{\delta\boldsymbol{B}} = \begin{bmatrix} \delta a \\ \delta b \end{bmatrix}$,$\underset{n,1}{\boldsymbol{W}} = [-a_0 - b_0x_1 + y_1, -a_0 - b_0x_2 + y_2, \cdots, -a_0 - b_0x_n + y_n]^{\mathrm{T}}$。

在最小二乘准则 $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \min$ 下,得到回归方程的未知参数估计值:

$$\delta\boldsymbol{B} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W} \tag{4}$$

参数估值为:

$$a = a_0 + \delta a, b = b_0 + \delta b \tag{5}$$

以 y 作 x 的一元线性回归模型为:

$$x = c + dy \tag{6}$$

应用上述同样估计方法可以得到回归参数 c 和 d 的估计值。

如文献[2]中所述,大量的计算实践证明,y 关于 x、x 关于 y 的回归是不一致的,即两个回归方程的参数不能满足条件 $a+bc=0$,$bd-1=0$。

1.2　顾及自变量误差的条件平差算法

为了解决这种不一致性,文献[3]提出了同时顾及 x 和 y 误差的条件平差法。文献[3]的推导结果为:

$$\delta\boldsymbol{B} = [\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}})^{-1}\boldsymbol{A}]^{-1}\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}})^{-1}\boldsymbol{W} \tag{7}$$

因为

$$\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}} = b_0^2\boldsymbol{I} + \boldsymbol{I} = (b_0^2 + 1)\boldsymbol{I} \tag{8}$$

将式(8)代入式(7),整理后和式(4)一样,所以其解算结果必然完全一致。因此,对于直线回归方程式 $y=a+bx$ 而言,顾及 x 自变量误差的解算和不考虑其误差影响的解算应用上述条件平差法对回归参数 a、b 的估值解算结果无影响,所建立的 y 关于 x 的回归方程和普通最小二乘回归完全相同。

如果用回归方程式 $x=c+dy$ 按照上述条件平差解算方法,其解算的估值 c、d 也和普通最小二乘回归相同,因此,文献[3]中所述的方法实际上并未同时顾及 x 和 y 的误差,也没有解决一元线性回归的不一致性问题。

1.3　考虑自变量误差的间接平差算法

文献[4]所建立的模型的解算结果为:

$$\begin{bmatrix} \delta\boldsymbol{x} \\ \delta\boldsymbol{B} \end{bmatrix} = \begin{bmatrix} (1+b_0^2)\boldsymbol{I} & b_0\boldsymbol{A} \\ \boldsymbol{A}^{\mathrm{T}}b_0 & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} \end{bmatrix}^{-1}\begin{bmatrix} \boldsymbol{W}_1 + b_0\boldsymbol{W}_2 \\ \boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}_2 \end{bmatrix} \tag{9}$$

由分块矩阵求逆有:

$$\begin{bmatrix} (1+b_0^2)\boldsymbol{I} & b_0\boldsymbol{A} \\ \boldsymbol{A}^{\mathrm{T}}b_0 & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} \end{bmatrix}^{-1} = \begin{bmatrix} \dfrac{1}{1+b_0^2}\boldsymbol{I} + \dfrac{b_0^2}{1+b_0^2}\boldsymbol{A}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}} & -b_0\boldsymbol{A}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1} \\ -b_0(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}} & (1+b_0^2)(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1} \end{bmatrix} \tag{10}$$

将式(10)代入式(9)中,整理可得:

$$\delta\boldsymbol{B} = -b_0(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{W}_1 + b_0\boldsymbol{W}_2) + (1+b_0^2)(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}_2 = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W} \tag{11}$$

同时顾及取 $x_i^0 = x_i$,则式(11)和式(4)完全相同,所以其解算结果是一致的。

2　基于整体最小二乘法的一元线性回归解法

2.1　解算原理

为了同时顾及 x 和 y 的误差,并消除 y 关于 x、x 关于 y 两种回归方程解算结果的不一致性问题,建立如下直线模型:

$$a(x_i - \varepsilon_{x_i}) + b(y_i - \varepsilon_{y_i}) = 1, i = 1,2,\cdots,n \tag{12}$$

相应的条件方程式为:

$$a(x_i + v_{x_i}) + b(y_i + v_{y_i}) = 1 \tag{13}$$

取 a_0、b_0 为 a、b 的近似值,舍去二次项 $\delta a v_{x_i}$ 和 $\delta b v_{y_i}$ 的乘积项,则条件方程为:

$$a_0v_{x_i} + b_0v_{y_i} + x_i\delta a + y_i\delta b + a_0x_i + b_0y_i - 1 = 0 \tag{14}$$

矩阵形式为:

$$\underset{n,2n}{\boldsymbol{F}}\underset{2n,1}{\boldsymbol{V}} + \underset{n,2}{\boldsymbol{B}}\underset{2,1}{\delta\boldsymbol{B}} - \underset{n,1}{\boldsymbol{W}_3} = \underset{n,1}{0} \tag{15}$$

其中,

$$\boldsymbol{F} = \begin{bmatrix} a_0 & b_0 & & & & \\ & & a_0 & b_0 & & \\ & & & & \ddots & \\ & & & & a_0 & b_0 \end{bmatrix}, \boldsymbol{B} = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix},$$

$$\boldsymbol{W}_3 = \begin{bmatrix} -a_0x_1 - b_0y_1 + 1 \\ -a_0x_2 - b_0y_2 + 1 \\ \vdots \\ -a_0x_n - b_0y_n + 1 \end{bmatrix}$$

在最小二乘条件 $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{V} = \min$ 下,构造条件极值函数,可得[7]:

$$\boldsymbol{K} = -(\boldsymbol{F}\boldsymbol{P}^{-1}\boldsymbol{F}^{\mathrm{T}})^{-1}(\boldsymbol{B}\delta\boldsymbol{B} - \boldsymbol{W}_3) \tag{16}$$

$$\delta\boldsymbol{B} = [\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{F}\boldsymbol{P}^{-1}\boldsymbol{F}^{\mathrm{T}})^{-1}\boldsymbol{B}]^{-1}\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{F}\boldsymbol{P}^{-1}\boldsymbol{F}^{\mathrm{T}})^{-1}\boldsymbol{W}_3 \tag{17}$$

从而可以得到直线方程 $ax+by=1$ 相应的回归方程为：

$$y=\frac{1}{b}-\frac{a}{b}x \text{ 或 } x=\frac{1}{a}-\frac{b}{a}y \tag{18}$$

观测值 x、y 的改正数向量为：

$$\boldsymbol{V}=-\boldsymbol{P}^{-1}\boldsymbol{F}^{\mathrm{T}}(\boldsymbol{F}\boldsymbol{P}^{-1}\boldsymbol{F}^{\mathrm{T}})^{-1}(\boldsymbol{B}\delta\boldsymbol{B}-\boldsymbol{W}_3) \tag{19}$$

当 x_i 和 y_i 为独立等精度时，

$$\delta\boldsymbol{B}=[\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}})^{-1}\boldsymbol{B}]^{-1}\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}})^{-1}\boldsymbol{W}_3 \tag{20}$$

由于 $\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}=(a_0^2+b_0^2)\boldsymbol{I}$，所以整理上式得：

$$\delta\boldsymbol{B}=[\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}})^{-1}\boldsymbol{B}]^{-1}\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}})^{-1}\boldsymbol{W}_3=(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{W}_3 \tag{21}$$

$$\boldsymbol{V}=-\frac{1}{a_0^2+b_0^2}\boldsymbol{F}^{\mathrm{T}}(\boldsymbol{B}\delta\boldsymbol{B}-\boldsymbol{W}_3) \tag{22}$$

2.2 精度估计

观测数据的标准差估计公式为：

$$\hat{\sigma}=\sqrt{\boldsymbol{V}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{V}/f} \tag{23}$$

式中，f 为自由度。

按协因数传播律可得：

$$\boldsymbol{Q}_{\delta\boldsymbol{B}}=(\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{F}\boldsymbol{P}^{-1}\boldsymbol{F}^{\mathrm{T}})^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{F}\boldsymbol{P}^{-1}\boldsymbol{F}^{\mathrm{T}})^{-1}\boldsymbol{Q}_{\boldsymbol{W}_3\boldsymbol{W}_3}\cdot(\boldsymbol{F}\boldsymbol{P}^{-1}\boldsymbol{F}^{\mathrm{T}})^{-1}\boldsymbol{B}(\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{F}\boldsymbol{P}^{-1}\boldsymbol{F}^{\mathrm{T}})^{-1}\boldsymbol{B})^{-1}=(\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{F}\boldsymbol{P}^{-1}\boldsymbol{F}^{\mathrm{T}})^{-1}\boldsymbol{B})^{-1}$$

当 x_i 和 y_i 为独立等精度时，

$$\boldsymbol{Q}_{\delta\boldsymbol{B}}=(a_0^2+b_0^2)(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}$$

令 $(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}=\begin{bmatrix}Q_{11} & Q_{12}\\ Q_{12} & Q_{22}\end{bmatrix}$，则可得 a、b 的方差为：

$$D(a)=\hat{\sigma}^2(a_0^2+b_0^2)Q_{11}=\hat{\sigma}^2Q_a$$
$$D(b)=\hat{\sigma}^2(a_0^2+b_0^2)Q_{22}=\hat{\sigma}^2Q_b$$
$$D(a,b)=\hat{\sigma}^2(a_0^2+b_0^2)Q_{12}=\hat{\sigma}^2Q_{ab} \tag{24}$$

3 计算实例

以文献[8]例5-2的数据为样本观测值，共计25个点，设回归直线方程为：

$$y=a+bx \text{ 或 } x=c+dy$$

表1给出了所有样本点的数据。表2为几种方法得出的结果比较。

表1 观测样本值

Tab.1 The Value of the observations

编号	y	x	编号	y	x
1	10.98	35.3	14	9.57	39.1
2	11.13	29.7	15	10.94	46.8
3	12.51	30.8	16	9.58	48.5
4	8.40	58.8	17	10.09	59.3
5	9.27	61.4	18	8.11	70.0
6	8.73	71.3	19	6.83	70.0
7	6.36	74.4	20	8.88	74.5
8	8.50	76.6	21	7.68	72.1
9	7.82	70.7	22	8.47	58.1
10	9.14	57.5	23	8.86	44.6
11	8.24	46.4	24	10.38	33.4
12	12.19	28.9	25	11.08	28.6
13	11.88	28.1			

表2 几种方法比较

Tab.2 Comparisons of the Results

	普通回归模型法		Golub算法		本文新模型
因变量	y	x	y	x	
自变量	x	y	x	y	
模型	$y=a+bx$	$x=c+dy$	$y=a+bx$	$x=c+dy$	$ax+by=1$
模型解算结果	$y=13.628\,4-0.079\,9x$	$x=136.993\,7-8.952\,5y$	$y=14.195\,2-0.089\,7x$	$x=158.271\,1-11.149\,6y$	$0.006\,32x+0.070\,43y=1$
统一表达	$y=13.628\,4-0.079\,9x$	$y=15.302\,2-0.111\,7x$	$y=14.195\,2-0.089\,7x$	$y=14.195\,2-0.089\,7x$	$y=14.198\,1-0.089\,7x$

从表2可以看出，普通最小二乘法的解算结果会因自变量选择的不同而不同，但整体最小二乘法和新模型解算方法的解算结果是一致的。

4 结 语

针对普通回归中的不一致性问题，本文给出了能同时顾及 x、y 误差的整体解算模型和新解算模型，并将模型的建立和解法纳入了测量平差的理论体系。随着空间数据采集方法的多样性、多源性，本文所提出的方法对于解决目前遇到的许多新问题具有一定的参考价值，对进一步研究同时考虑函数模型中系数阵误差的整体解算具有一定的启发，为以后更深入地研究整体最小二乘在测量数据处理中的应用打下了研究基础。

参 考 文 献

[1] 丁勇. 直线回归的最小面积法[J]. 工程数学学报，2003，20(3)：47-50

[2] 李雄军. 对 X 和 Y 方向最小二乘线性回归的讨论[J]. 计量技术，2005(1)：50-52

[3] 王安怡，陶本藻. 顾及自变量误差的回归分析理论和方法[J]. 勘测科学技术，2005(3)：29-32

[4] 刘大杰，史文中，童小华，等. GIS空间数据的精度分析与质量控制[M]. 上海：上海科学技术文献出版社，1999

[5] 孟晓林,刘大杰,朱照宏. 道路曲线数字化数据处理的平差模型[J]. 同济大学学报,1998,26(6):669-673

[6] 童小华,刘大杰. 道路曲线数字化数据的联合平差模型[J]. 武汉大学学报·信息科学版,2001,26(1):64-69

[7] 武汉大学测量平差学科组编. 误差理论与测量平差基础[M]. 武汉:武汉大学出版社,2003

[8] 腾素珍,冯敬海. 数理统计学[M]. 大连:大连理工大学出版社,2005

第一作者简介:鲁铁定,副教授,博士生。主要研究方向为测绘数据处理理论和大地测量。

E-mail:tdlu@ecit.edu.cn

Modeling and Algorithm of Linear Regression Based on Total Least Squares

LU Tieding[1,2]　*TAO Benzao*[1]　*ZHOU Shijian*[2,3]

(1 School of Geodesy and Geomatics, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

(2 School of Geoscience and Surveying Engineering, East China Institute of Technology, 56 Xuefu Road, Fuzhou 344000, China)

(3 Jiangxi Academy of Sciences, Shangfang Road, Nanchang 330029, China)

Abstract: The independent variables error and variables error are further studied. A new regression model is given, and calculation algorithms are deduced. Numerical experiments are used to demonstrate correctness of the new method, and the results are same as those of the total least square method. The algorithms proposed in this paper simplify the complicated matrix decomposing calculation, bring TLS into category of surveying adjustment.

Key words: linear regression; total least squares; surveying adjustment

About the first author: LU Tieding, associate professor, Ph. D candidate. His research orientation is surveying data processing and geodesy. E-mail: tdlu@ecit.edu.cn

(上接第 471 页)

Dam Deformation Prediction Based on Wavelet Transform and Support Vector Machine

WANG Xinzhou[1]　*FAN Qian*[1,2]　*XU Chengquan*[1,2]　*LI Zhao*[1,2]

(1 School of Geodesy and Geomatics, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

(2 Research Center for Hazard Monitoring and Prevention, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

Abstract: A novel model based on wavelet transform and support vector machine for dam deformation prediction is presented. Firstly, through the wavelet transform, deformation time series is decomposed into different frequency components. Then, according to the different characteristics of the decomposed components, different support vector machines are constructed to forecast the components. Finally, the predicted results of the components are reconstructed to be used as the final prediction result of deformation. The calculation result shows that this model has higher forecasting precision and greater generality ability.

Key words: wavelet transform; support vector machine; dam deformation prediction

About the first author: WANG Xinzhou, Ph. D, professor, Ph. D supervisor, majors in the theory and application of spatial data processing. E-mail: fanqian1981@163.com