

文章编号:1671-5942(2014)05-0110-04

考虑自变量误差的线性回归迭代算法^{* 1}

汪奇生¹⁾ 杨德宏¹⁾ 杨根新²⁾

(1) 昆明理工大学国土资源工程学院, 昆明 650093

(2) 云南国土资源职业学院, 昆明 650093

摘要 为解决线性回归中自变量含误差的问题, 根据间接平差模型和总体最小二乘原理推导了一种总体最小二乘迭代算法。算例验证了该算法的有效性和可行性。

关键词 总体最小二乘; 线性回归; 自变量; 迭代算法; 平差模型

中图分类号:P207

文献标识码:A

ITERATION ALGORITHM OF LINEAR REGRESSION CONSIDERING THE ERROR OF INDEPENDENT VARIABLES

Wang Qisheng¹⁾, Yang Dehong¹⁾ and Yang Genxin²⁾

(1) Faculty of Land Resource Engineering, KUST, Kunming 650217
(2) Yunnan Land and Resources Vocational College, Kunming 650217

Abstract Considering the error of adjustment problem for independent variable in linear regression, an iteration algorithm of total least squares was derived according to indirect adjustment model and total least squares theory. The algorithm is simple and easy to programming. The result indicates that the algorithms is more effective and more feasible than other algorithms.

Key words: total least squares; linear regression; independent variable; iteration algorithm; adjustment model

线性拟合通常是采用最小二乘法确定回归参数, 在自变量不含误差的前提下得到最优参数值。实际上, 拟合数据往往都含有偶然误差, 需要在平差的同时考虑线性回归中自变量的误差。文献[1-2]提出以正交距离残差平方和极小为准则的正交最小二乘法, 实质上是一个总体最小二乘^[3]问题。由于总体最小二乘的常規矩阵分解算法不利于测量数据的处理, 发展了一些迭代算法^[4-8], 将系数矩阵中的全部元素当成含有误差来处理, 这就不适宜线性回归参数的求解。一般来讲, 求解线性回归的总体最小二乘解采用混合总体最小二乘法^[9], 但解算复杂, 难以理解且没有考虑到测量平差的优势。本

文根据线性回归模型的特点推导了一种迭代算法, 结果与混合总体最小二乘法一致。

1 总体最小二乘模型及迭代算法

1.1 总体最小二乘模型

线性回归数学模型为:

$$y = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \cdots + \hat{a}_{n-1} x_{n-1} \quad (1)$$

当有多组观测值并考虑自变量的误差时, 其总体最小二乘平差模型为:

$$\mathbf{Y} + \mathbf{e} = (\mathbf{B} + \mathbf{E}_B) \boldsymbol{\beta} \quad (2)$$

* 收稿日期: 2013-11-20

作者简介: 汪奇生, 男, 1989 年生, 硕士研究生, 主要研究方向为大地测量数据处理。E-mail: wangqisheng0702@163.com。

$$\text{式中, } \mathbf{B} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1(n-1)} \\ 1 & x_{21} & \cdots & x_{2(n-1)} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{m1} & \cdots & x_{m(n-1)} \end{bmatrix}, \mathbf{\hat{B}} = \begin{bmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \vdots \\ \hat{a}_n \end{bmatrix}, \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

\mathbf{e} 为 \mathbf{Y} 的误差向量, \mathbf{E}_B 为 \mathbf{B} 的误差矩阵。考虑自变量的误差时, 其系数矩阵也含有误差, 但同时含有常数列和误差项, 一般采用混合总体最小二乘法^[9] 进行解算, 而不是常规的总体最小二乘迭代算法, 因为迭代算法将系数矩阵 \mathbf{B} 的全部元素都进行了改正, 而不能固定常数列。

1.2 算法原理

将式(1)进行等价转换:

$$\begin{aligned} b_0 y + b_1 x_1 + b_2 x_2 + \cdots + b_{n-1} x_{n-1} &= 1, b_0 = \frac{1}{\hat{a}_0}, b_{n-1} = \\ -\frac{\hat{a}_{n-1}}{\hat{a}_0}, n &= 1, 2, \dots \end{aligned} \quad (3)$$

当有多组观测值并考虑自变量的误差时, 式(3)可以表示为:

$$(\mathbf{A} + \mathbf{E}_A) \mathbf{X} = \mathbf{L} \quad (4)$$

$$\text{式中, } \mathbf{A} = \begin{bmatrix} y_1 & x_{11} & \cdots & x_{1(n-1)} \\ y_2 & x_{21} & \cdots & x_{2(n-1)} \\ \vdots & \vdots & \cdots & \vdots \\ y_m & x_{m1} & \cdots & x_{m(n-1)} \end{bmatrix},$$

$$\mathbf{E}_A = \begin{bmatrix} v_{y_1} & v_{x_{11}} & \cdots & v_{x_{1(n-1)}} \\ v_{y_2} & v_{x_{21}} & \cdots & v_{x_{2(n-1)}} \\ \vdots & \vdots & \cdots & \vdots \\ v_{y_m} & v_{x_{m1}} & \cdots & v_{x_{m(n-1)}} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \vdots \\ \hat{b}_n \end{bmatrix}, \mathbf{L} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

式(4)还可以表示为:

$$\begin{aligned} \mathbf{AX} - \mathbf{L} &= -\mathbf{E}_A \mathbf{X} = -(X^T \otimes \mathbf{I}_m) \text{vec}(\mathbf{E}_A) = \\ &\quad -(X^T \otimes \mathbf{I}_m) \mathbf{V} \end{aligned} \quad (5)$$

式中, \otimes 为矩阵的克罗内克积, $\text{vec}(\mathbf{E}_A)$ 是将矩阵 \mathbf{E}_A 按列从左到右拉直得到的列向量化矩阵。 \mathbf{V} 是平差模型中 $(m \times n) \times 1$ 的误差向量, $\mathbf{V} = \text{vec}(\mathbf{E}_A)$ 。若自变量与因变量独立等精度, 根据总体最小二乘原理, 相应的误差期望和方差为:

$$\mathbf{V} = \text{vec}(\mathbf{E}_A) \sim [0, \sigma_0^2 \cdot (\mathbf{I}_m \otimes \mathbf{I}_n)] \quad (6)$$

式中, \mathbf{I}_m 和 \mathbf{I}_n 分别为 m 和 n 阶单位矩阵。由式(6), $\mathbf{Q}_v = \mathbf{I}_{m \times n}, \mathbf{I}_{m \times n}$ 为 $m \times n$ 阶单位矩阵。相应的平差准则为:

$$\mathbf{V}^T \mathbf{V} = \min \quad (7)$$

考虑到 $(X^T \otimes \mathbf{I}_m)(X \otimes \mathbf{I}_m) = X^T X$ 为一常数, 设 $\mathbf{V} = (X \otimes \mathbf{I}_m) \mathbf{K}$, 代入式(5)可得:

$$\mathbf{K} = -(\mathbf{AX} - \mathbf{L}) / (X^T X) \quad (8)$$

由式(8)再根据 $\mathbf{V} = (X \otimes \mathbf{I}_m) \mathbf{K}$ 可知:

$$\mathbf{V}^T \mathbf{V} = (\mathbf{AX} - \mathbf{L})^T (\mathbf{AX} - \mathbf{L}) / X^T X \quad (9)$$

另将式(5)表示为:

$$\bar{\mathbf{V}} = -\mathbf{E}_A \mathbf{X} = \mathbf{AX} - \mathbf{L} \quad (10)$$

根据式(5)将 $\bar{\mathbf{V}}$ 表达为:

$$\bar{\mathbf{V}} = -\mathbf{E}_A \mathbf{X} = (X^T \otimes \mathbf{I}_m) \text{vec}(\mathbf{E}_A) = (X^T \otimes \mathbf{I}_m) \mathbf{V} \quad (11)$$

则根据协因数传播定律可得:

$$\mathbf{Q}_V = (X^T \otimes \mathbf{I}_m) \mathbf{Q}_V (X \otimes \mathbf{I}_m) = (X^T X) \mathbf{I}_m \quad (12)$$

由式(10)、(12)可得:

$$\mathbf{V}^T \mathbf{Q}_V^{-1} \bar{\mathbf{V}} = (\mathbf{AX} - \mathbf{L})^T (\mathbf{AX} - \mathbf{L}) / (X^T X) \quad (13)$$

由此可知 $\bar{\mathbf{V}}^T \mathbf{Q}_V^{-1} \bar{\mathbf{V}} = \mathbf{V}^T \mathbf{V}$, 则式(7)的平差准则等价为:

$$\bar{\mathbf{V}}^T \mathbf{Q}_V^{-1} \bar{\mathbf{V}} = \min \quad (14)$$

按照拉格朗日原理求解目标函数的自由极值:

$$\frac{\partial (\bar{\mathbf{V}}^T \mathbf{Q}_V^{-1} \bar{\mathbf{V}})}{\partial \mathbf{X}} = 0 \quad (15)$$

将式(15)求偏导并顾及式(11)、(13)得:

$$\begin{aligned} \frac{\partial (\bar{\mathbf{V}}^T \mathbf{Q}_V^{-1} \bar{\mathbf{V}})}{\partial \mathbf{X}} &= \frac{\partial (\bar{\mathbf{V}}^T \mathbf{Q}_V^{-1} \bar{\mathbf{V}})}{\partial \bar{\mathbf{V}}} \cdot \frac{\partial \bar{\mathbf{V}}}{\partial \mathbf{X}} + \frac{\partial (\bar{\mathbf{V}}^T \mathbf{Q}_V^{-1} \bar{\mathbf{V}})}{\partial \mathbf{Q}_V^{-1}} \cdot \\ &\quad \frac{\partial \mathbf{Q}_V^{-1}}{\partial \mathbf{X}} = 2\bar{\mathbf{V}}^T \mathbf{Q}_V^{-1} \mathbf{A} - \frac{2\bar{\mathbf{V}}^T \mathbf{V} X^T}{(X^T X)^2} = 0 \end{aligned} \quad (16)$$

将式(16)转置、整理得:

$$\mathbf{A}^T \mathbf{Q}_V^{-1} \bar{\mathbf{V}} - \frac{\mathbf{X} \bar{\mathbf{V}}^T \mathbf{V}}{(X^T X)^2} = 0 \quad (17)$$

将式(11)、(12)代入式(17), 整理可得:

$$\mathbf{A}^T (\mathbf{AX} - \mathbf{L}) / X^T X - \mathbf{X} (\mathbf{AX} - \mathbf{L})^T (\mathbf{AX} - \mathbf{L}) / X^T X^2 = 0 \quad (18)$$

设 $\mathbf{v} = (\mathbf{AX} - \mathbf{L})^T (\mathbf{AX} - \mathbf{L}) / X^T X$, 整理得:

$$\mathbf{A}^T (\mathbf{AX} - \mathbf{L}) = \mathbf{X} (\mathbf{AX} - \mathbf{L})^T (\mathbf{AX} - \mathbf{L}) / X^T X = \mathbf{X}_v \quad (19)$$

对式(19)整理得参数估值:

$$\mathbf{X} = (\mathbf{A}^T \mathbf{A})^{-1} (\mathbf{A}^T \mathbf{L} + \mathbf{X}_v) \quad (20)$$

其单位权中误差为:

$$\sigma_0^2 = (\mathbf{V}^T \mathbf{V}) / (m - n) \quad (21)$$

根据式(11)、(13)可知, $\mathbf{V}^T \mathbf{V} = \bar{\mathbf{V}}^T \mathbf{Q}_V^{-1} \bar{\mathbf{V}} = (\mathbf{AX} - \mathbf{L})^T (\mathbf{AX} - \mathbf{L}) / X^T X = \mathbf{v}$ 。

1.3 解算步骤

1) 由式(1)根据最小二乘原理求得回归参数估值 a_0, a_1, a_n , 再根据式(3)将其变换为 b_0, b_1, b_n , 并组成回归参数的初值 $\mathbf{X}^{(0)} = [b_0 \ b_1 \ \cdots \ b_n]^T$ 。

2) 按下式计算新的回归参数值:

$$\begin{aligned} \mathbf{v}^{(i)} &= (\mathbf{AX}^{(i)} - \mathbf{L})^T (\mathbf{AX}^{(i)} - \mathbf{L}) / (X^{(i)T} X^{(i)}) \\ \mathbf{X}^{(i+1)} &= (\mathbf{A}^T \mathbf{A})^{-1} (\mathbf{A}^T \mathbf{L} + \mathbf{X}^{(i)} \mathbf{v}^{(i)}) \end{aligned} \quad (22)$$

3) 重复步骤 2), 当 $|X^{(i+1)} - X^{(i)}| < \varepsilon$ 时停止迭代。

4) 输出参数估值, 按式(21)求单位权中误差。

通过上述迭代方法即可得到式(3)的方程 $\hat{b}_0y + \hat{b}_1x_1 + \hat{b}_2x_2 + \cdots + \hat{b}_{n-1}x_{n-1} = 1$, 相应的线性回归方程为:

$$y = \hat{a}_0 + \hat{a}_1x_1 + \hat{a}_2x_2 + \cdots + \hat{a}_{n-1}x_{n-1} \quad (23)$$

式中, $\hat{a}_0 = \frac{1}{\hat{b}_0}$, $\hat{a}_{n-1} = -\frac{\hat{b}_{n-1}}{\hat{b}_0}$, $n = 1, 2, \dots$

2 实例分析

为验证本文算法的合理性, 利用文献[10]中的观测数据(表1)拟合 $y = a + bx$, 其自变量和因变量都含有误差。分别设计如下5种方案对交换自变量与因变量分别进行拟合。

表1 观测数据
Tab. 1 Observation data

序号	1	2	3	4	5	6	7	8	9	10	11	12
x	2	2	3	4	5	4.5	5.5	7.5	8	9	10	11
y	30	35	40	45	50	55	66	75	85	100	110	120

方案1: 按最小二乘方法, 以 x 为自变量、 y 为因变量拟合线性方程 $y = a + bx$, 得到的结果为 $y = 9.5077 + 9.7470x$; 以 y 为自变量、 x 为因变量拟合线性方程 $x = c + dy$, 得到的结果为 $x = -0.8246 + 0.1004y$, 将其转换得 $y = 8.2160 + 9.9638x$ 。

方案2: 按总体最小二乘的奇异值分解法, 以 x 为自变量、 y 为因变量拟合线性方程 $y = a + bx$, 得到的结果为 $y = 15.0902 + 9.0111x$; 以 y 为自变量、 x 为因变量拟合线性方程 $x = c + dy$, 得到的结果为 $x = -1.6746 + 0.1110y$, 将其转换得 $y = 15.0902 + 9.0111x$ 。

方案3: 按文献[5]的总体最小二乘迭代算法, 以 x 为自变量、 y 为因变量拟合线性方程 $y = a + bx$, 得到的结果为 $y = 15.0902 + 9.0111x$; 以 y 为自变量、 x 为因变量拟合线性方程 $x = c + dy$, 得到的结果为 $x = -1.6746 + 0.1110y$, 将其转换得 $y = 15.0902 + 9.0111x$ 。

方案4: 按文献[9]的混合总体最小二乘法, 以 x 为自变量、 y 为因变量来拟合线性方程 $y = a + bx$, 得到的结果为 $y = 8.2291 + 9.9615x$; 以 y 为自变量、 x 为因变量来拟合线性方程 $x = c + dy$, 得到的结果为 $x = -0.8261 + 0.1004y$, 将其转换得 $y = 8.2291 + 9.9615x$ 。

方案5: 按本文算法以 x 为自变量、 y 为因变量拟合线性方程 $y = a + bx$, 得到的结果为 $y = 8.2291 + 9.9615x$; 以 y 为自变量、 x 为因变量来拟合线性方程 $x = c + dy$, 得到的结果为 $x = -0.8261 + 0.1004y$, 将其转换得 $y = 8.2291 + 9.9615x$ 。

比较5种方案可以看出, 对于自变量含误差的

线性回归问题, 方案1采用的最小二乘法并没有考虑自变量的误差, 因此交换自变量和因变量拟合出来的结果不一致。其他4种方案交换自变量和因变量拟合出来的结果一致, 但4种方案拟合出两套结果。方案2和方案3的结果相同, 方案4和方案5的结果相同。这是因为, 方案2和方案3采用的总体最小二乘奇异值分解法和迭代解法都考虑到了线性回归平差模型中系数矩阵 B 的误差, 但将系数矩阵所有元素都看成是含误差的, 并对其常数列也进行了改正。交换自变量和因变量拟合出来的结果虽然一致, 但实质上其拟合结果是有偏差的。而方案4采用混合总体最小二乘来解算, 既考虑了自变量的误差又顾及了平差模型中系数矩阵 B 的常数列。只改正系数矩阵中含有误差的元素, 得到的结果是合理的。方案5采用本文的算法, 将回归模型进行等价转换, 其系数矩阵由自变量和因变量组成, 都含有误差, 从而将原平差模型系数矩阵 B 的常数列转换成常数向量, 得到的结果和方案4相同。

本文算法将线性回归模型进行等价转换, 系数矩阵由自变量和因变量组成, 平差模型的误差也只存在于系数矩阵内, 常数列向量不含误差。但本文将模型 $y = \hat{a}_0 + \hat{a}_1x_1 + \hat{a}_2x_2 + \cdots + \hat{a}_nx_n$ 转换为 $\hat{b}_0y + \hat{b}_1x_1 + \hat{b}_2x_2 + \cdots + \hat{b}_nx_n = 1$ 是基于回归模型中存在常数项 \hat{a}_0 , 如果不存在常数项就不能进行这样的转换, 其总体最小二乘解算也可以采用奇异值分解法或常规的迭代算法。另外, 如果不进行等价转换, 就不能采用本文算法。因为常规的平差模型中还有观测向量的误差, 而本文通过转换后将观测向量(即因变量)的误差归入到系数矩阵中, 原平差模型系数矩阵的常数列转换为常数向量, 故本文给出的针对线性回归的总体最小二乘迭代算法是有效可行的。

3 结语

1) 对于自变量含误差的线性回归问题, 采用总体最小二乘法求得的结果更合理。但不宜采用常规的迭代算法, 因为常规的迭代算法不能顾及系数矩阵的常数列而是将系数矩阵所有元素都当成含有误差来处理, 这对线性回归是不合理的。

2) 本文给出的迭代算法是针对自变量含误差的线性回归问题, 既能充分考虑线性回归中自变量的误差, 又能顾及平差模型中系数矩阵的常数列, 得到的结果与混合总体最小二乘相同。而与混合总体最小二乘相比, 本文算法充分考虑了测量平差的优势, 推导过程简单且更适于程序实现。

参 考 文 献

- 1 丁克良,欧吉坤,赵春梅. 正交最小二乘曲线拟合法[J]. 测绘科学,2007,32(3):18-19. (Ding Keliang, Ou Jikun, Zhao Chunmei. Methods of the least-squares orthogonal distance fitting[J]. Science of Surveying and Mapping, 2007, 32 (3):18-19)
- 2 丁克良,刘全利,陈翔. 正交距离圆曲线拟合方法[J]. 测绘科学,2008,33(10):72-73. (Ding Keliang, Liu Quanli, Chen Xiang. Fitting of circles based on orthogonal distance [J]. Science of Surveying and Mapping, 2008, 33(10):72 -73)
- 3 Golub G H, Vanl C F. An analysis of the total least squares problem[J]. Siam J Numer Anal, 1980, 17:883-893.
- 4 鲁铁定,周世健. 总体最小二乘的迭代解法[J]. 武汉大学学报:信息科学版,2010,35(11):1 351-1 354. (Lu Tieding, Zhou Shijian. An iteration for the total least squares estimation[J]. Geomatics and Information Science of Wuhan University, 2010, 35(11):1 351-1 354)
- 5 孔建,姚宜斌,吴寒. 整体最小二乘的迭代解法[J]. 武汉大学学报:信息科学版,2010,35(6):711-714. (Kong Jian, Yao Yibin, Wu Han. Iterative method for total least-squares[J]. Geomatics and Information Science of Wuhan University, 2010, 35(6):711-714)
- 6 许超铃. 基于整体最小二乘的参数估计新方法及精度评定[J]. 测绘通报,2011(10):1-4. (Xu Chaoqian. New method of parameters estimation and accuracy evaluation based on TLS[J]. Bulletin of Surveying and Mapping, 2011 (10):1-4)
- 7 邱卫宁,齐公玉,田丰瑞. 整体最小二乘求解线性模型的改进算法[J]. 武汉大学学报:信息科学版,2010,35(6):708-710. (Qiu Weining, Qi Gongyu, Tian Fengrui. An improved algorithm of total least squares for linear models[J]. Geomatics and Information Science of Wuhan University, 2010, 35(6):708-710)
- 8 邱卫宁. 测量数据处理理论与方法[M]. 武汉:武汉大学出版社,2008. (Qiu Weining. The theory and method of surveying data processing [M]. Wuhan: Wuhan University Press, 2008)
- 9 丁克良,沈云中,欧吉坤. 整体最小二乘法直线拟合[J]. 辽宁工程技术大学学报:自然科学版,2010,29(1):44-47. (Ding Keliang, Sheng Yunzhon, Ou Jikun. Methods of line-fitting based on total least-squares[J]. Journal of Liaoning Technical University: Natural Science, 2010, 29(1):44 -47)
- 10 何晓群. 实用回归分析[M]. 北京:高等教育出版社,2008. (He Xiaoqun. Practical regression analysis[M]. Beijing: Higher Education Press, 2008)

(上接第 109 页)

- 7 Dong D. Spatiotemporal filtering using principal component analysis and Karhunen-Loeve expansion approaches for regional GPS network analysis[J]. Journal of Geophysical Research, 2006, 111(B3):3 405-3 421.
- 8 Williams S D P. Error analysis of continuous GPS position time series[J]. Journal of Geophysical Research, 2004, 109 (B3):412-430.
- 9 Williams S D P. CATS: GPS coordinate time series analysis software[J]. GPS Solutions, 2008, 12(2):147-153.
- 10 黄立人. GPS 基准站时间序列的噪声特性分析[J]. 大地测量与地球动力学, 2006(2):31-33. (Huang Liren. Noise properties in time series of coordinate component at GPS fiducial stations[J]. Journal of Geodesy and Geodynamics, 2006(2):31-33)
- 11 Wdowinski S. Southern California permanent GPS geodetic array: Spatial filtering of daily positions for estimating coseismic and postseismic displacements induced by the 1992 Landers earthquake[J]. J Geophys Res, 1997, 102(B8):18 057-18 070.
- 12 胡守超,伍吉仓,孙亚峰. 区域 GPS 网三种时空滤波方法的比较[J]. 大地测量与地球动力学, 2009(3):95-99. (Hu Shouchao, Wu Jicang, Sun Yafeng. Comparison among three spatiotemporal filtering methods for regional GPS networks analysis[J]. Journal of Geodesy and Geodynamics, 2009(3):95-99)
- 13 Agnew D C. The time-domain behavior of power-law noises [J]. Geophys Res Lett, 1992, 19(4):333-336.
- 14 姜卫平. 利用连续 GPS 观测数据分析水库长期变形[J]. 测绘学报, 2012, 41(5):682-689. (Jiang Weiping. Analysis of long-term deformation of reservoir using continuous GPS observations[J]. Acta Geodaetica et Cartographica Sinica, 2012, 41(5):682-689)
- 15 蒋志浩. 顾及有色噪声影响的 CGCS2000 下我国 CORS 站速度估计[J]. 测绘学报, 2010, 39(4):355-363. (Jiang Zihao. Velocity estimation on the colored noise properties of CORS network in China based on the CGCS2000 frame [J]. Acta Geodaetica et Cartographica Sinica, 2010, 39(4):355-363)