Models and Algorithms of Linear Regression Based on Total least squares

Ding Shijun[1, 2] **Jiang Weiping[3]** **Shen Zhijuan[1]**

1) School of Geodesy and Geomation, Wuhan University, Wuhan, China

2) Key Laboratory of Precise Eengineering and Industry Surveying, State Bureau of Surveying and Mapping, Wuhan, China

3) GNSS Engineering Research Center of Wuhan University, Wuhan, China

Abstract： In classical regression analysis, Error of independent variable x is ususlly not taken into account during regression analysis。When The independent variable X and dependent variable y are with errors, from adjustment model, solution methods are derived from the models of the condition adjustment and indirect adjustment based on the total least squares principle, and the equivalence of the two kinds of solution methods is proved in theory. Finally, some conclusions are drawn.

Key Words： Total least squares ; Regression analysis; Adjustment model; Equivalence

1 Introduction

Total least squares (TLS) was firstly proposed by Golub and Van Loan[1], during the last decade, a lot of theoretical studies have been done in TLS, such as the algorithm of TLS, conditions of solution and the relations between TLS and Least Squares[2]. Some practical problems, for example, signal processing, statistical calculation, regression analysis, can be mapped into the problem of TLS. In the field of mapping and survey, regression analysis is one popular method of measurement data processing, and the traditional solution of the model is to get the best estimates of regression parameters based on least square principle and by assuming independent variable x is without errors and dependent variable y is observations with random errors. Considering a group of measurement data $(x_i, y_i), i = 1,2,\cdots,n$, if the errors of measurement data x_i and y_i are both taken into account, the solutions of regression parameters can be summed up as the problem of TLS. Some researches have been presented in Reference [1-6], two different methods are discussed respectively in Ref. [3], and Ref. [4] analyzes and compares the problems in Ref. [3], but fails to give reasonable explanations, resulting in biased conclusions. Ref. [5] proves the equivalence of solutions of total least squares linear regression in condition adjustment and indirect adjustment, but lacks the proof for the equivalence of precision estimation. Therefore, this paper, adopting the methods of data processing in adjustment of measurement, does an in-depth study for the solutions of TLS linear regression parameters, aiming at laying a foundation for TLS theories in the application of measurement data processing.

2 Linear regression model of independent variables without errors

For the convenience of discussion, take single regression as an example. Suppose measurement point (x_i, y_i), then the unary linear regression model will be

$$y_i = a + bx_i + \Delta_i \qquad (i = 1,2,\cdots,n) \tag{1}$$

Where a, b are regression parameters, Δ_i is the true error of measurement y_i.

Given independent variable x_i is error free, let the approximate values of unknown parameters a, b are a_0, b_0, and their corrections are da, db .The error equation according to indirect adjustment is,

$$v_{y_i} = da + x_i db - l_i \tag{2}$$

Where $l_i = -a_0 - x_i b_0 + y_i$, represented by matrix,

$$\underset{n\times 1}{\mathbf{V}} = \underset{n\times 2}{\mathbf{A}}\, \underset{2\times 1}{d\mathbf{B}} - \underset{n\times 1}{\mathbf{L}} \tag{3}$$

Where $\underset{n\times 2}{\mathbf{A}} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_3 \end{bmatrix}$, $\underset{2\times 1}{d\mathbf{B}} = \begin{bmatrix} da \\ db \end{bmatrix}$, $\underset{n\times 1}{\mathbf{L}} = \begin{bmatrix} -a_0 - b_0 x_1 + y_1 \\ -a_0 - b_0 x_2 + y_2 \\ \vdots \\ -a_0 - b_0 x_n + y_n \end{bmatrix} = \begin{bmatrix} l_1 \\ l_2 \\ \vdots \\ l_n \end{bmatrix}$

Suppose $\mathbf{P}_{yy}$ is the weight matrix for measurement y, based on Least Squares principle $\mathbf{V}^T \mathbf{P}_{yy} \mathbf{V} = \min$, then the estimates of regression parameter corrections are

$$d\mathbf{B} = (\mathbf{A}^T \mathbf{P}_{yy} \mathbf{A})^{-1} A^T \mathbf{P}_{yy} \mathbf{L} \tag{4}$$

3 Linear regression model of independent variables with errors

3.1 Condition adjustment model of independent variable with errors

Given independent variable x is with errors and independent from y, the regression model equation is

$$y_i + v_{y_i} = \hat{a} + \hat{b}(x_i + v_{x_i}) \tag{5}$$

Where v_{x_i}, v_{y_i} are corrections of observations. Apparently the equation above contains the second order terms of parameters and measurement corrections, so substitute unknown parameters a, b with approximations a_0, b_0 in Equ. (5), linearize and leave out the second order terms, the equation can be reduced to the following,

$$-v_{y_i}+b_0v_{x_i}+da+x_idb+a_0+b_0x_i-y_i=0 \qquad (6)$$

Matrix expression is as follows,

$$\underset{n\times 2n}{\mathbf{E}}\,\underset{2n\times 1}{\mathbf{V}}+\underset{n\times 2}{\mathbf{A}}\,\underset{2\times 1}{d\mathbf{B}}-\underset{n\times 1}{\mathbf{L}}=\underset{n\times 1}{\mathbf{0}} \qquad (7)$$

Where $\underset{n\times 2n}{\mathbf{E}}=\left[b_0\underset{n\times n}{\mathbf{I}} \quad -\underset{n\times n}{\mathbf{I}}\right]^T$, $\mathbf{I}$ is unit matrix, $\underset{2n\times 1}{\mathbf{V}}=\left[\mathbf{V}_x^T \quad \mathbf{V}_y^T\right]^T$

$\mathbf{V}_x=\left[v_{x_1} \quad v_{x_2} \quad \cdots \quad v_{x_n}\right]^T$, $\mathbf{V}_y=\left[v_{y_1} \quad v_{y_2} \quad \cdots \quad v_{y_n}\right]^T$, $\underset{2\times 1}{d\mathbf{B}}=\left[da \quad db\right]^T$

$\underset{n\times 1}{\mathbf{L}}=\left[-a_0-b_0x_1+y_1 \quad -a_0-b_0x_2+y_2 \quad \cdots \quad -a_0-b_0x_n+y_n\right]^T$

Equ. (7) is the constraint equation with residuals and parameter corrections. One equation can be derived from each measurement point, thus we have n equations with $2n+2$ undetermined values (2 parameters a, b, $2n$ v_x and v_y), Least Square principle is used here.

Given independent variable x and dependent variable y are independent and with different precision, the random model is

$$\mathbf{P}=\begin{bmatrix}\mathbf{P}_{xx} & 0\\ 0 & \mathbf{P}_{yy}\end{bmatrix} \qquad (8)$$

Based on least square principle $\mathbf{V}^T\mathbf{P}\mathbf{V}=\min$, the extremal function is formed as

$$\varphi=\mathbf{V}^T\mathbf{P}\mathbf{V}-2\mathbf{K}^T(\mathbf{E}\mathbf{V}+\mathbf{A}d\mathbf{B}-\mathbf{L})$$

$\mathbf{K}$ is the matrix of correlates. The partial derivatives of $\boldsymbol{V}$, $d\boldsymbol{B}$ from the extremal function are

$$\frac{\partial\varphi}{\partial\mathbf{V}}=0, \quad \mathbf{V}=\mathbf{P}^{-1}\mathbf{E}^T\mathbf{K} \qquad (9)$$

$$\frac{\partial\varphi}{\partial d\mathbf{B}}=0, \quad \mathbf{A}^T\mathbf{K}=0 \qquad (10)$$

Combining (7)、(9) and (10), the unique solution of $\mathbf{K}$ 、 $\mathbf{V}$ and $d\mathbf{B}$ can be acquired

$$\mathbf{K}=-(\mathbf{E}\mathbf{P}^{-1}\mathbf{E}^T)^{-1}(\mathbf{A}d\mathbf{B}-\mathbf{L}) \qquad (11)$$

$$d\mathbf{B}=[\mathbf{A}^T(\mathbf{E}\mathbf{P}^{-1}\mathbf{E}^T)^{-1}\mathbf{A}]^{-1}\mathbf{A}^T(\mathbf{E}\mathbf{P}^{-1}\mathbf{E}^T)^{-1}\mathbf{L} \qquad (12)$$

$$\mathbf{V}=-\mathbf{P}^{-1}\mathbf{E}^T(\mathbf{E}\mathbf{P}^{-1}\mathbf{E}^T)^{-1}(\mathbf{A}d\mathbf{B}-\mathbf{L}) \qquad (13)$$

For $\mathbf{E}\mathbf{P}^{-1}\mathbf{E}^T=(b_0\mathbf{I}-\mathbf{I})\begin{bmatrix}\mathbf{P}_{xx} & 0\\ 0 & \mathbf{P}_{yy}\end{bmatrix}^{-1}(b_0\mathbf{I}-\mathbf{I})^T=b_0^2\mathbf{P}_{xx}^{-1}+\mathbf{P}_{yy}^{-1}$, Equ. (12)、(13) can be shown as

$$d\mathbf{B} = [\mathbf{A}^T(b_0^2\mathbf{P}_{xx}^{-1} + \mathbf{P}_{yy}^{-1})^{-1}\mathbf{A}]^{-1}\mathbf{A}^T(b_0^2\mathbf{P}_{xx}^{-1} + \mathbf{P}_{yy}^{-1})^{-1}\mathbf{L} \quad (14)$$

$$\mathbf{V} = -\begin{bmatrix} b_0\mathbf{P}_{xx}^{-1} \\ -\mathbf{P}_{yy}^{-1} \end{bmatrix}(b_0^2\mathbf{P}_{xx}^{-1} + \mathbf{P}_{yy}^{-1})^{-1}(\mathbf{A}d\mathbf{B} - \mathbf{L}) \quad (15)$$

And the estimates of regression parameters are $\hat{a} = a_0 + da$, $\hat{b} = b_0 + db$.

3.2 Indirect adjustment model of independent variable with errors

Suppose independent variable $\hat{x}_i$ and parameters $\hat{a}$, $\hat{b}$ are unknown, then the number of unknown parameters is $n+2$, x_i , y_i are measurements, and their number is $2n$. Let the approximate values of parameters are x_{i0} , a_0 and b_0 , their corrections are dx_i , da and db respectively, the adjustment equation is given by

$$\hat{x}_i = x_i + v_{x_i}$$
$$\hat{y}_i = y_i + v_{y_i} = \hat{a} + \hat{b}\hat{x}_i$$

Let $x_{i0} = x_i$, substitute the approximations of unknown parameters into the equation, linearize and leave out the second order terms, the error equation is reduced to

$$\begin{cases} v_{x_i} = dx_i \\ v_{y_i} = (b_0 \quad 1 \quad x_i)\begin{bmatrix} dx_i \\ da \\ db \end{bmatrix} + a_0 + b_0 x_i - y_i \end{cases} \quad (16)$$

Matrix expression is shown as

$$\mathbf{V} = \mathbf{C}d\mathbf{Z} - \mathbf{L}_1 \quad (17)$$

where

$$\mathbf{C} = \begin{bmatrix} \underset{n\times n}{\mathbf{I}} & \underset{n\times 2}{\mathbf{0}} \\ b_0\underset{n\times n}{\mathbf{I}} & \underset{n\times 2}{\mathbf{A}} \end{bmatrix}, \ d\mathbf{Z} = \begin{bmatrix} d\mathbf{X}^T & d\mathbf{B}^T \end{bmatrix}^T, \ d\mathbf{X} = \begin{bmatrix} dx_1 & dx_2 & \cdots & dx_n \end{bmatrix}^T, \ d\mathbf{B} = \begin{bmatrix} da & db \end{bmatrix}^T$$

$$\mathbf{L}_1 = \begin{bmatrix} \underset{n\times 1}{\mathbf{0}} & \underset{n\times 1}{\mathbf{L}}^T \end{bmatrix}^T, \ \mathbf{V} = \begin{bmatrix} \mathbf{V}_x^T & \mathbf{V}_y^T \end{bmatrix}^T$$

$$\mathbf{V}_x = \begin{bmatrix} v_{x_1} & v_{x_2} & \cdots & v_{x_n} \end{bmatrix}^T, \ \mathbf{V}_y = \begin{bmatrix} v_{y_1} & v_{y_2} & \cdots & v_{y_n} \end{bmatrix}^T$$

Based on indirect adjustment principle, the normal equation is formed as

$$\begin{bmatrix} \mathbf{P}_{xx} + b_0^2\mathbf{P}_{yy} & b_0\mathbf{P}_{yy}\mathbf{A} \\ b_0\mathbf{A}^T\mathbf{P}_{yy} & \mathbf{A}^T P_{yy}\mathbf{A} \end{bmatrix}\begin{bmatrix} d\mathbf{X} \\ d\mathbf{B} \end{bmatrix} = \begin{bmatrix} b_0\mathbf{P}_{yy}\mathbf{L} \\ \mathbf{A}^T\mathbf{P}_{yy}\mathbf{L} \end{bmatrix} \quad (18)$$

From the first equation of Equ.(18), we get

$$d\mathbf{X} = -b_0(\mathbf{P}_{xx} + b_0^2\mathbf{P}_{yy})^{-1}\mathbf{P}_{yy}(\mathbf{A}d\mathbf{B} - \mathbf{L}) \tag{19}$$

Substitute the equation above into the second equation of (18), we obtain

$$d\mathbf{B} = (\mathbf{A}^T\mathbf{M}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{M}\mathbf{L} \tag{20}$$

Where $\mathbf{M} = \mathbf{P}_{yy} - b_0^2\mathbf{P}_{yy}(\mathbf{P}_{xx} + b_0^2\mathbf{P}_{yy})^{-1}\mathbf{P}_{yy}$

3.3 Equivalence of the two adjustment models of independent variables with errors

To prove the equivalence of the two adjustment methods, the Equ.(21) is needed to be proved after comparing Equ. (20) and Equ. (14),

$$\mathbf{M} = (b_0^2\mathbf{P}_{xx}^{-1} + \mathbf{P}_{yy}^{-1})^{-1} \tag{21}$$

For $\mathbf{M} = \mathbf{P}_{yy} - b_0^2\mathbf{P}_{yy}(\mathbf{P}_{xx} + b_0^2\mathbf{P}_{yy})^{-1}\mathbf{P}_{yy}$, using matrix inversion formula, one can get

$$\begin{aligned}(\mathbf{P}_{xx} + b_0^2\mathbf{P}_{yy})^{-1} &= \frac{1}{b_0^2}\mathbf{P}_{yy}^{-1} - \frac{1}{b_0^2}\mathbf{P}_{yy}^{-1}(\mathbf{P}_{xx}^{-1} + \frac{1}{b_0^2}\mathbf{P}_{yy}^{-1})^{-1}\frac{1}{b_0^2}\mathbf{P}_{yy}^{-1} \\ &= \frac{1}{b_0^2}(\mathbf{P}_{yy}^{-1} - \mathbf{P}_{yy}^{-1}(b_0^2\mathbf{P}_{xx}^{-1} + \mathbf{P}_{yy}^{-1})^{-1}\mathbf{P}_{yy}^{-1})\end{aligned} \tag{22}$$

Substitute Equ.(22), the equation can be deduced to,

$$\mathbf{M} = \mathbf{P}_{yy} - b_0^2\mathbf{P}_{yy}(\mathbf{P}_{xx} + b_0^2\mathbf{P}_{yy})^{-1}\mathbf{P}_{yy} = (b_0^2\mathbf{P}_{xx}^{-1} + \mathbf{P}_{yy}^{-1})^{-1} \tag{23}$$

Thus Equ. (21) is proved, and the solutions expressed by Equ. (14) is equivalent to that by Equ. (20).

Similarly, substitute Equ. (19) into Equ. (17), we get

$$\begin{aligned}\mathbf{V} &= \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ b_0\mathbf{I} & \mathbf{A} \end{bmatrix} \begin{bmatrix} -b_0(\mathbf{P}_{xx} + b_0^2\mathbf{P}_{yy})^{-1}\mathbf{P}_{yy}(\mathbf{A}d\mathbf{B} - \mathbf{L}) \\ d\mathbf{B} \end{bmatrix} - \begin{bmatrix} \mathbf{0} \\ \mathbf{L} \end{bmatrix} \\ &= \begin{bmatrix} -b_0(\mathbf{P}_{xx} + b_0^2\mathbf{P}_{yy})^{-1}\mathbf{P}_{yy}(\mathbf{A}d\mathbf{B} - \mathbf{L}) \\ -b_0^2(\mathbf{P}_{xx} + b_0^2\mathbf{P}_{yy})^{-1}\mathbf{P}_{yy}(\mathbf{A}d\mathbf{B} - \mathbf{L}) + \mathbf{A}d\mathbf{B} - \mathbf{L} \end{bmatrix}\end{aligned} \tag{24}$$

and Equ. (24) can be reorganized as

$$\mathbf{V} = \begin{bmatrix} -b_0(\mathbf{P}_{xx} + b_0^2\mathbf{P}_{yy})^{-1}\mathbf{P}_{yy} \\ -(b_0^2(\mathbf{P}_{xx} + b_0^2\mathbf{P}_{yy})^{-1}\mathbf{P}_{yy} - \mathbf{I}) \end{bmatrix}(\mathbf{A}d\mathbf{B} - \mathbf{L}) \tag{25}$$

Theorem. if matrix $\mathbf{A}$ and $\mathbf{B}$ are regular, then $(\mathbf{A} + \mathbf{B})^{-1} = \mathbf{A}^{-1}(\mathbf{A}^{-1} + \mathbf{B}^{-1})\mathbf{B}^{-1}$. According to the theorem above, the first equation of Equ. (25) can be reduced to

$$-b_0(\mathbf{P}_{xx}+b_0^2\mathbf{P}_{yy})^{-1}\mathbf{P}_{yy}=-b_0\mathbf{P}_{xx}^{-1}(b_0^2\mathbf{P}_{xx}^{-1}+\mathbf{P}_{yy}^{-1})^{-1} \tag{26}$$

The second equation of Equ. (25) can be reduced to the following based on inversion formula of matrice

$$-(b_0^2(\mathbf{P}_{xx}+b_0^2\mathbf{P}_{yy})^{-1}\mathbf{P}_{yy}-\mathbf{I})=\mathbf{P}_{yy}^{-1}(b_0^2\mathbf{P}_{xx}^{-1}+\mathbf{P}_{yy}^{-1}) \tag{27}$$

Substitute Equ. (26), (27) into Equ. (25) to get the equation below

$$\mathbf{V}=-\begin{bmatrix} b_0\mathbf{P}_{xx}^{-1} \\ -\mathbf{P}_{yy}^{-1}\end{bmatrix}(b_0^2\mathbf{P}_{xx}^{-1}+\mathbf{P}_{yy}^{-1})^{-1}(\mathbf{A}d\mathbf{B}-\mathbf{L}) \tag{28}$$

After comparision between Equ. (28) and (14), the conclusion that the corrections from the two methods are equal is proved.

In particular, if independent variable and dependent variable are independent from each other and with equal precision, then $\mathbf{M}=\mathbf{P}_{yy}/(b_0^2+1)$, and the Equ.(20) can be reduced to $d\mathbf{B}=(\mathbf{A}^T\mathbf{P}_{yy}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{P}_{yy}\mathbf{L}$, which means the values of regression parameters calculated from Equ. (20), (14), (4) are equal. In such a case, the linear regression solutions of independent variable without errors and with errors are the same, but from Equ. (28), (3), we know that the corrections of two cases are different, which means their precisions are different.

3.4 Linear regression solution precision estimate of independent variable with errors

The standard deviation is estimated as the equation below

$$\hat{\sigma}=\sqrt{\frac{\mathbf{V}^T\mathbf{P}\mathbf{V}}{n-t}}=\sqrt{\frac{\mathbf{V}_x^T\mathbf{P}_{xx}\mathbf{V}_x+\mathbf{V}_y^T\mathbf{P}_{yy}\mathbf{V}_y}{n-t}} \tag{29}$$

Where n is the number of observation points, t is the number of regression parameters, and as with unary linear regression, $t=2$.

According to variance propagation law, the variance resulted from Equ. (20) is shown as

$$\mathbf{D}_{d\mathbf{B}}=(\mathbf{A}^T\mathbf{M}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{M}\mathbf{D}_{\mathbf{L}}\mathbf{M}\mathbf{A}^T(\mathbf{A}^T\mathbf{M}\mathbf{A})^{-1} \tag{30}$$

Where $\mathbf{D}_{\mathbf{L}}$ is the variance of $\mathbf{L}$. Due to $l_i=-a_0-x_ib_0+y_i$, its vector form is

$$\underset{n\times 1}{\mathbf{L}}=\begin{pmatrix}-b_0\underset{n\times n}{\mathbf{I}} & \underset{n\times n}{\mathbf{I}}\end{pmatrix}\begin{pmatrix}\underset{n\times 1}{\mathbf{X}} \\ \underset{n\times 1}{\mathbf{Y}}\end{pmatrix}-a_0e=\mathbf{E}\begin{pmatrix}\underset{n\times 1}{\mathbf{X}} \\ \underset{n\times 1}{\mathbf{Y}}\end{pmatrix}-a_0\mathbf{e} \tag{31}$$

Where $\mathbf{e}=[1 \quad 1 \quad \cdots \quad 1]^T$. From Equ. (31), the variance matrix of $\mathbf{L}$ is

$$\mathbf{D_L} = \hat{\sigma}^2 \mathbf{E}\mathbf{P}^{-1}\mathbf{E}^T = \hat{\sigma}^2 \mathbf{M}^{-1} \tag{32}$$

Substitute Equ. (32) into Equ. (30), then get

$$\mathbf{D}_{d\mathbf{B}} = \hat{\sigma}^2 (\mathbf{A}^T \mathbf{M}\mathbf{A})^{-1} \tag{33}$$

4 Conclusion

Solutions of total least square linear regression are derived from the models of the condition adjustment and indirect adjustmen when the independent variable x and dependent variable y are with errors. Adopting the models of the condition adjustment and indirect adjustment in adjustment of measurement, the equivalence of solutions of the parameters and the equivalence of precision estimation are proved based on Total Least Square in thory. Finally precision estimate equations of solution are given for total least square linear regression. algorithms of the single linear regression based on total least squares are also applied to multiple linear regression.

AKNOWLEDGEMENT

We would like to thank Professors Tao Benzao for several helpful conversations during the preparation of this manuscript, and are grateful to the support of the National Nature Science Foundation of China (No.41174009).

Reference

1 Golub G H, Van Loan C F.An anlysis of the total least squares problem.SIAM J. Numer.Anal., 1980, 17(6): 883-893

2 Van Huffel S.,and J.Vandewalle, The Total Least Squares Problem: Computational Aspects and Analysis, SIAM, Philadelphia,1991

3 Wan Anyi, Tao BenZao.Theory and Method of Regression Analysis for Error of Independent Variable.Technology of Exploration and survey Science, 2005(3):29-32

4 Lu Tieding, et al. Linear regression modeling and solution based on Total Least Squares. Geomatics and Information Science of Wuhan University, 2008, 33(5):504-507

5 Zhou Shijian, Lu Tieding, The Equvalence of the Calculating Methodology for Bi-variable Linear Regression. Jiangxi Science, 2009, 6: 867-870

6 Kong Jian,Yao Yibin,et al. Solving Coordinate Tansfomtion Parameters Based on Total Least Squares Regression, Journal of Geogesy and Geodynamics,2010(3): 74-78

7 Qiu Weining,et al.The theory and method of surveying data processing.wuhan:Wuhan University Press,2008