Article ID: 1003 - 6326(2000)02 - 0280 - 04

Data perturbation analysis of a linear model①

JIN Feng-xiang(靳奉祥)[1], Michel Mayoud[2],
LU Xiu-shan(卢秀山)[1], Jean-Pierre Quesnel[2]
1. Geoscience Deptartment, Shandong University of Science and Technology, Taian 271019, P. R. China;
2. Positioning Metrology and Surveying Group, European Organization for Nuclear Research (CERN), Geneva, Switzerland

Abstract: The linear model features were carefully studied in the cases of data perturbation and mean shift perturbation. Some important features were also proved mathematically. The results show that the mean shift perturbation is equivalent to the data perturbation, that is, adding a parameter to an observation equation means that this set of data is deleted from the data set. The estimate of this parameter is its predicted residual in fact.

Key words: linear model; data perturbation; mean shift perturbation **Document code:** A

1 INTRODUCTION

As one of the most important basic models, linear model has been great widely used in each and every field. Studies on it are more and more deep and detailed. Rao totally analyzed its parameter's estimates and developed linear model estimation theory[1,2]. Seber introduced the model parameter's selecting method[3]. A great contribution in this field was also made by Chen and Wang[4,5]. And some others also did some research works[6]. In addition many specialists in geodesy also made a big contribution to develop the theory and its applications in surveying data processing. Koch worked on the Bayes estimation of variance component[7~9]. Zhou studied its robust features and robust estimation, Chen developed this theory and used it for deformation data processing and deformation analysis[10]. The authors studied its statistic features and influential features[11~13].

Statistic diagnostics theory is just one important branch of them. It takes the linear model as the object to be studied, and takes the relationships between the data and model, data and model parameter, parameter and model, as the main content to follow. The theory describes carefully the internal relationship between data and model, data and parameter. It becomes a very effective tool for data analysis and model analysis. It is also very practical for information getting from surveying data. Many scholars have been deeply studying the model[14]. The authors have been also doing some research works in this field[12,13,15]. In this paper some features of a linear model are studied in the cases of data perturbation. Work done is in the field of data analysis and model analysis theory. Some most important conclusions are proved theoretically. The results presented here are the basic theory of surveying data processing and analysis.

2 PERTURBATION MODEL AND ESTIMATING

There is a linear model as follows:

$$\underset{n\times 1}{Y} = \underset{n\times p}{X}\cdot\underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon} \qquad \varepsilon \to N(0,\ \sigma^2\Omega^{-1}) \tag{1}$$

Divide (X, Y) into two sets (X_J, Y_J) and $(X(J), Y(J))$; X_J and Y_J are k dimensional vectors; $X(J)$ and $Y(J)$ are $n-k$ dimensional vectors. $\Omega=\begin{bmatrix}\omega(J) & 0\\ 0 & \omega_J\end{bmatrix}$, where $\omega(J)$ and ω_J being the diagonal matrix.

In correspondence with the formula above, Eqn.(1) could be written as

$$\begin{bmatrix}Y(J)\\ Y_J\end{bmatrix} = \begin{bmatrix}X(J)\\ X_J\end{bmatrix}\beta + \begin{bmatrix}\varepsilon(J)\\ \varepsilon_J\end{bmatrix} \qquad \begin{bmatrix}\varepsilon(J)\\ \varepsilon_J\end{bmatrix} \to N\left(0,\ \sigma^2\begin{bmatrix}\omega^{-1}(J) & 0\\ 0 & \omega_J^{-1}\end{bmatrix}\right) \tag{2}$$

2.1 Perturbation model of data deletion

After deleting the data subset (X_J, Y_J), Eqn.(2) becomes into

$$Y(J) = X(J)\beta + \varepsilon(J) \qquad \varepsilon(J) \to N(0,\ \sigma^2\omega^{-1}(J)) \tag{3}$$

① **Foundation item:** Project Y98E09080 supported by the Natural Science Foundation of Shandong Province and project supported by European Organization for Nuclear Research **Received date:** Apr.13, 1999; **accepted date:** Jan.7, 2000

Writing $\hat{\beta}$, $\hat{\beta}(J)$ as the least square parameter estimation of Eqns.(1) and (3) respectively, and the sum of residuals squares are written as RSS and $RSS(J)$ respectively. Then the following formulas would be tenable

$$\hat{\beta}(J) = \hat{\beta} - N^{-1} X_J^{\mathrm{T}} \omega_J (I - J_J)^{-1} \hat{e}_J \tag{4}$$

$$RSS = RSS(J) + \hat{e}_J^{\mathrm{T}} \omega_J (I - J_J)^{-1} \hat{e}_J \tag{5}$$

where

$$\hat{\beta} = N^{-1} X^{\mathrm{T}} \Omega Y;$$

$$N = X^{\mathrm{T}} \Omega X;$$

$$J_J = X_J N^{-1} X_J^{\mathrm{T}} \omega_J;$$

$$\hat{e}_J = Y_J - \hat{Y}_J = Y_J - X_J \hat{\beta}.$$

And there are some features in Eqn.(5) as follows:

$$RSS \to \sigma^2 \chi^2(n - p) \tag{6}$$

$$RSS(J) \to \sigma^2 \chi^2(n - p - k) \tag{7}$$

$$\hat{e}_J^{\mathrm{T}} \omega_J (I - J_J)^{-1} \hat{e}_J \to \sigma^2 \chi^2(k) \tag{8}$$

Here, Eqns.(7) and (8) are dependant each other, at the same time

$$\frac{RSS(J)}{RSS} \to \beta\left(\frac{n - p - k}{2}, \frac{k}{2}\right) \tag{9}$$

$$\frac{RSS - RSS(J)}{RSS} = \frac{r_J^2}{n - p} \to \beta\left(\frac{k}{2}, \frac{n - p - k}{2}\right) \tag{10}$$

$$\frac{RSS - RSS(J)}{RSS(J)} = \frac{\hat{e}_J^{\mathrm{T}} \omega_J (I - J_J)^{-1} \hat{e}_J}{RSS(J)} \cdot \frac{n - p - k}{k} \to F(k, n - p - k) \tag{11}$$

Eqns.(4), (5), (8), (9) and (10) are proved next.

Prove Ⅰ Eqn.(4) would be proved as:

From Eqns.(1) and (3) the least square estimator of the model parameter β are

$$\hat{\beta} = N^{-1} X^{\mathrm{T}} \Omega Y \tag{12}$$

$$\hat{\beta}(J) = N^{-1}(J) X^{\mathrm{T}}(J) \omega(J) Y(J) \tag{13}$$

where

$$N(J) = X^{\mathrm{T}}(J) \omega(J) X(J)$$

$$\begin{aligned} N^{-1}(J) &= [X^{\mathrm{T}}(J) \omega(J) X(J) + X_J^{\mathrm{T}} \omega_J X_J - X_J^{\mathrm{T}} \omega_J X_J]^{-1} \\ &= N^{-1} + N^{-1} X_J^{\mathrm{T}} (\omega_J^{-1} - X_J N^{-1} X_J^{\mathrm{T}})^{-1} X_J N^{-1} \end{aligned} \tag{14}$$

$$X^{\mathrm{T}}(J) \omega(J) Y(J) = X^{\mathrm{T}} \Omega Y - X_J^{\mathrm{T}} \omega_J Y_J \tag{15}$$

Bring Eqns.(14) and (15) into (13), we have

$$\hat{\beta}(J) = \hat{\beta} + N^{-1} X_J^{\mathrm{T}} (\omega_J^{-1} - X_J N^{-1} X_J^{\mathrm{T}})^{-1} \hat{Y}_J - N^{-1} X_J^{\mathrm{T}} (\omega_J + (\omega_J^{-1} - X_J N^{-1} X_J^{\mathrm{T}})^{-1} X_J N^{-1} X_J^{\mathrm{T}} \omega_J) Y_J \tag{16}$$

where $\hat{Y}_J$ is the estimator of Y_J with Eqn.(1), that is $\hat{Y}_J = X_J N^{-1} X^{\mathrm{T}} \Omega Y$.

Because

$$(I - A)^{-1} = I + (I + A)^{-1} A \tag{17}$$

is tenable.

Therefore

$$\begin{aligned} \hat{\beta}(J) &= \hat{\beta} + N^{-1} X_J^{\mathrm{T}} (\omega_J^{-1} - X_J N^{-1} X_J^{\mathrm{T}})^{-1} \hat{Y}_J - N^{-1} X_J^{\mathrm{T}} \omega_J (I - X_J N^{-1} X_J^{\mathrm{T}} \omega_J)^{-1} Y_J \\ &= \hat{\beta} - N^{-1} X_J^{\mathrm{T}} \omega_J (I - J_J)^{-1} \hat{e}_J \end{aligned} \tag{18}$$

Thus Eqn.(4) has been proved.

Prove Ⅱ Eqn.(5) would be proved as:

From the definition, it is known

$$RSS = (Y - X\hat{\beta})^{\mathrm{T}} \Omega (Y - X\hat{\beta}) \tag{19}$$

Put Eqn.(4) into that above, we get

$$\begin{aligned} RSS = &(Y - X\hat{\beta}(J) - XN^{-1} X_J^{\mathrm{T}} (\omega_J^{-1} - X_J N^{-1} X_J^{\mathrm{T}})^{-1} \hat{e}_J)^{\mathrm{T}} \Omega \cdot \\ &(Y - X\hat{\beta}(J) - XN^{-1} X_J^{\mathrm{T}} (\omega_J^{-1} - X_J N^{-1} X_J^{\mathrm{T}})^{-1} \hat{e}_J) \end{aligned} \tag{20}$$

Bring it into blocks, let

$$(\omega_J^{-1} - X_J N^{-1} X_J^{\mathrm{T}})^{-1} = B$$

we have

$$RSS = \begin{bmatrix} \hat{e}(J) - X(J) N^{-1} X_J^{\mathrm{T}} B \hat{e}_J \\ \hat{e}_J \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \omega(J) & 0 \\ 0 & \omega_J \end{bmatrix} \begin{bmatrix} \hat{e}(J) - X(J) N^{-1} X_J^{\mathrm{T}} B \hat{e}_J \\ \hat{e}_J \end{bmatrix} \tag{21}$$

where $\hat{e}(J)$ is the residual vector of $Y(J)$ with Eqn.(3), $\hat{e}_J$ is the residual vector of Y_J with Eqn.(1). So

we have

$$RSS = \hat{e}^{\mathrm{T}}(J)\,\omega(J)\,\hat{e}(J) - \hat{e}^{\mathrm{T}}(J)\,\omega(J)\,X(J)\,N^{-1}\,X_J^{\mathrm{T}}B\hat{e}_J - \hat{e}_J^{\mathrm{T}}BX_JN^{-1}\,X^{\mathrm{T}}(J)\,\omega(J)\,\hat{e}(J) + \hat{e}_J^{\mathrm{T}}BX_JN^{-1}\,X^{\mathrm{T}}(J)\,\omega(J)\,X(J)\,N^{-1}\,X_J^{\mathrm{T}}B\hat{e}_J + \hat{e}_J^{\mathrm{T}}\omega_J\hat{e}_J \quad (22)$$

From the linear model theory, it is well known

$$X^{\mathrm{T}}(J)\,\omega(J)\,\hat{e}(J) = 0 \quad (23)$$

$$RSS(J) = \hat{e}^{\mathrm{T}}(J)\,\omega(J)\,\hat{e}(J) \quad (24)$$

$$X^{\mathrm{T}}(J)\,\omega(J)\,X(J) = N - X_J^{\mathrm{T}}\omega_J X_J \quad (25)$$

Put Eqns.(23), (24) and (25) into (22), we get

$$RSS = RSS(J) + \hat{e}_J^{\mathrm{T}}BX_JN^{-1}\,X_J^{\mathrm{T}}\omega_J(\omega_J^{-1} - X_JN^{-1}\,X_J^{\mathrm{T}})\,B\hat{e}_J + \hat{e}_J^{\mathrm{T}}\omega_J\hat{e}_J \quad (26)$$

$$RSS = RSS(J) + \hat{e}_J^{\mathrm{T}}(\omega_J - BX_JN^{-1}\,X_J^{\mathrm{T}}\omega_J)\,\hat{e}_J \quad (27)$$

Because of

$$\begin{aligned}(\omega_J - BX_JN^{-1}\,X_J^{\mathrm{T}}\omega_J) &= \omega_J + (\omega_J^{-1} - X_JN^{-1}\,X_J^{\mathrm{T}})^{-1}\,X_JN^{-1}\,X_J^{\mathrm{T}}\omega_J \\ &= \omega_J(I + (I - X_JN^{-1}\,X_J^{\mathrm{T}}\omega_J)\,X_JN^{-1}\,X_J^{\mathrm{T}}\omega) = \omega_J(I - J_J)^{-1}\end{aligned} \quad (28)$$

Put Eqn.(28) into (27), Eqn.(5) is obtained.

Prove Ⅲ Eqn.(8) would be proved as:

From the theorem of the distribution of a quadratic form, if the two formulas are tenable

$$\hat{e}_J^{\mathrm{T}}\omega_J(I - J_J)^{-1}\hat{e}_J = Y^{T}Q_1\,Y \quad (29)$$

$$(Q_1\,\Omega^{-1})^2 = Q_1\,\Omega^{-1} \quad (30)$$

Then Eqn.(8) is tenable yet. The next step is to prove Eqns.(29) and (30) tenable. Suppose, a matrix

$$D = (d_1,\ d_2 \cdots d_k) \quad (31)$$

where d_i ($i = 1, 2, \cdots, k$) is a vector of n-dimensional, its $(n - k + i)$th is 1, others are 0, and

$$\left.\begin{aligned} &X_J = D^{\mathrm{T}}X,\ Y_J = D^{\mathrm{T}}Y, \\ &\hat{e}_J = D^{\mathrm{T}}\hat{e},\ \omega_J^{-1} = D^{\mathrm{T}}\,\Omega^{-1}\,D \end{aligned}\right\} \quad (32)$$

so

$$\hat{e}_J^{\mathrm{T}}(\omega_J^{-1} - X_JN^{-1}\,X_J^{\mathrm{T}})^{-1}\hat{e}_J = Y^{\mathrm{T}}Q^{\mathrm{T}}D(D^{\mathrm{T}}\,\Omega^{-1}\,D - D^{\mathrm{T}}XN^{-1}\,X^{\mathrm{T}}D)^{-1}\,D^{\mathrm{T}}QY \quad (33)$$

where $Q = I - XN^{-1}\,X^{\mathrm{T}}\,\Omega$, compared with Eqn.(29), we have

$$Q_1 = Q^{\mathrm{T}}D(D^{\mathrm{T}}\,\Omega^{-1}\,D - D^{\mathrm{T}}XN^{-1}\,X^{\mathrm{T}}D)^{-1}\,D^{\mathrm{T}}Q = Q^{\mathrm{T}}D(D^{\mathrm{T}}Q\Omega^{-1}\,D)^{-1}\,D^{\mathrm{T}}Q \quad (34)$$

$$\begin{aligned}(Q_1\,\Omega^{-1})^2 &= Q^{\mathrm{T}}D(D^{\mathrm{T}}Q\Omega^{-1}\,D)^{-1}\,D^{\mathrm{T}}Q\Omega^{-1}\,Q^{\mathrm{T}}D(D^{\mathrm{T}}Q\Omega^{-1}\,D)^{-1}\,D^{\mathrm{T}}Q\Omega^{-1} \\ &= Q^{\mathrm{T}}D(D^{\mathrm{T}}Q\Omega^{-1}\,D)^{-1}\,D^{\mathrm{T}}\,\Omega^{-1} = Q_1\,\Omega^{-1}\end{aligned} \quad (35)$$

Eqns.(29) and (30) are tenable. So Eqn.(8) is proved.

Prove Ⅳ Eqns.(7) and (8) are independent

Because of

$$RSS(J) = RSS - Y^{\mathrm{T}}Q_1\,Y = Y^{\mathrm{T}}(Q^{\mathrm{T}}\,\Omega Q - Q_1)\,Y \quad (36)$$

Therefore

$$(Q^{\mathrm{T}}\,\Omega Q - Q_1)\,\Omega^{-1}\,Q_1 = Q^{\mathrm{T}}\,\Omega Q\Omega^{-1}\,Q_1 - Q_1\,\Omega^{-1}\,Q_1 \quad (37)$$

$$Q_1\,\Omega^{-1}\,Q_1 = Q^{\mathrm{T}}D(D^{\mathrm{T}}Q\Omega^{-1}\,D)^{-1}\,D^{\mathrm{T}}Q\Omega^{-1}\,Q^{\mathrm{T}}D(D^{\mathrm{T}}Q\Omega^{-1}\,D)^{-1}\,D^{\mathrm{T}}Q = Q_1 \quad (38)$$

$$Q^{\mathrm{T}}\,\Omega Q\Omega^{-1}\,Q_1 = (QQ)^{\mathrm{T}}\,Q_1 = Q^{\mathrm{T}}\,Q_1 = (QQ)^{\mathrm{T}}D(D^{\mathrm{T}}Q\Omega^{-1}\,D)^{-1}\,D^{\mathrm{T}}Q = Q_1 \quad (39)$$

So Eqn.(37) equals to zero, Eqns.(7) and (8) are independent. And Eqns.(9) and (10) are tenable.

2.2 Perturbation model of mean shift

Change the model, Eqn.(1), into a model of mean shift

$$\begin{bmatrix} Y(J) \\ Y_J \end{bmatrix} = \begin{bmatrix} X(J) & 0 \\ X_J & I \end{bmatrix}\begin{bmatrix} \beta \\ \gamma \end{bmatrix} + \begin{bmatrix} \varepsilon(J) \\ \varepsilon_J \end{bmatrix} \quad (40)$$

Let $\hat{\beta}_a$, $\hat{\gamma}$ be the estimators of the parameters β, γ. The residual square sum can be written as RSS_a. The normal equation is

$$\begin{bmatrix} N & X_J^{T}\omega_J \\ \omega_J X_J & \omega_J \end{bmatrix}\begin{bmatrix} \hat{\beta}_a \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X^{\mathrm{T}}\,\Omega Y \\ \omega_J Y_J \end{bmatrix} \quad (41)$$

$$\begin{bmatrix} N & X_J^{\mathrm{T}}\omega_J \\ \omega_J X_J & \omega_J \end{bmatrix}^{-1} = \begin{bmatrix} N^{-1} + N^{-1}\,X_J^{\mathrm{T}}\omega_J(I - J_J)^{-1}\,X_JN^{-1} & -N^{-1}\,X_J^{\mathrm{T}}\omega_J(I - J_J)^{-1}\,\omega_J^{-1} \\ -(I - J_J)^{-1}\,X_JN^{-1} & (I - J_J)^{-1}\,\omega_J^{-1} \end{bmatrix} \quad (42)$$

The solutions are

$$\hat{\gamma} = (I - J_J)^{-1}\hat{e}_J \quad (43)$$

$$\hat{\beta}_a = \hat{\beta} - N^{-1}\,X_J^{\mathrm{T}}\omega_J\hat{\gamma} \quad (44)$$

That is

$$\hat{\beta}_a = \hat{\beta} - N^{-1} X_J^{\mathrm{T}} \omega_J (I - J_J)^{-1} \hat{e}_J \tag{45}$$

Comparing the formula above with Eqn.(4), we know

$$\hat{\beta}_a = \hat{\beta}(J) \tag{46}$$

The residual square sum is

$$RSS_a = \begin{bmatrix} Y(J) - X(J)\hat{\beta}_a \\ Y_J - X_J\hat{\beta}_a - \hat{\gamma} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \omega(J) & 0 \\ 0 & \omega_J \end{bmatrix} \begin{bmatrix} Y(J) - X(J)\hat{\beta}_a \\ Y_J - X_J\hat{\beta}_a - \hat{\gamma} \end{bmatrix} \tag{47}$$

where

$$Y_J - X_J\hat{\beta}_a - \hat{\gamma} = Y_J - X_J\hat{\beta} + X_J N^{-1} X_J^{\mathrm{T}} \omega_J \hat{\gamma} - \hat{\gamma} = \hat{e}_J - (I - J_J)\hat{\gamma} = \hat{e}_J - \hat{e}_J = 0 \tag{48}$$

Therefore

$$\begin{aligned} RSS_a &= [Y(J) - X(J)\hat{\beta}_a]^{\mathrm{T}} (\omega(J)) [Y(J) - X(J)\hat{\beta}_a] \\ &= [Y(J) - X(J)\hat{\beta}(J)]^{\mathrm{T}} (\omega(J)) [Y(J) - X(J)\hat{\beta}(J)] = RSS(J) \end{aligned} \tag{49}$$

Following the discussion above we could conclude that the model of data deletion and the model of mean shift are completely equivalent. Adding a mean shift parameter to an observational equation is equivalent to eliminate this observational equation, the estimator of the mean shift parameter is really the predicted residual of it.

3 CONCLUSIONS

From the perturbation analysis of a linear model some important conclusions could be gotten:

1) In the case that a linear model with a perturbation of data deletion, its estimators of parameters and residuals have the relations with those no perturbation, which are Eqns.(4) and (5), and have the statistic features expressed by Eqns.(6) to (11). Eqns.(7) and (8) are independent.

2) In the case that a linear model with a perturbation of mean shift, from Eqns.(46) and (49) we know that it is completely equivalent to that of the model of data deletion, that is, adding a mean shift parameter to an observational equation is equivalent to eliminate this observational equation, the estimator of the mean shift parameter is really the predicted residual of it.

REFERENCES

[1] Rao C R. Linear Statistical Inference and Its Applications (Second Edition) [M]. New York: John Wiley and Sons, 1973.

[2] WANG Song-gui. Linear Model Theory and Its Application [M], (in Chinese). Hefei: Anhui Educational Press, 1987.

[3] Seber G A F. Linear Regression Analysis [M]. New York: John Wiley and Sons, 1977.

[4] CHEN Xi-ru. Mathematical Statistics [M], (in Chinese). Beijing: Science Press, 1981.

[5] CHEN Xi-ru and WANG Song-gui. Modern Regression Analysis [M], (in Chinese). Hefei: Anhui Educational Press, 1987.

[6] ZHANG Jin-huai. Improvement and Rarameter Estimation of a Linear Model [M], (in Chinese). Changsha: Press of National University of Defence, 1992.

[7] Koch K R. Bayesian inference for variance components [J]. Manuscripta Geodactica, 1987, 12: 309 ~ 313.

[8] Koch K R. Bayesian statistics for variance components with informative and noninformative priors [J]. Manuscripta Geodactica, 1988, 13: 370 ~ 373.

[9] Koch K R. Bayesian Inference with Geodetic Applications [M]. Springer-Verlay, 1990.

[10] CHEN Yong-qi. Deformation Data Processing [M], (in Chinese). Beijing: Surveying and Mapping Press of China, 1988.

[11] JIN Feng-xiang and ZENG Z Q. Geometrical features and statistical test of GM model [J]. The Chinese Journal of Nonferrous Metals, 1996, 6(2): 1 ~ 6.

[12] JIN Feng-xiang. Influential Analysis Theory and Its Application on Deformation Monitoring [M], (in Chinese). Changsha: Central South University of Technology of China Press, 1994.

[13] JIN Feng-xiang and WANG Tong-xiao. Statistic Diagnostics and Deformation Monitoring Theory [M], (in Chinese). Beijing: Agricultural Science and Technology Publishing House of China Press, 1995.

[14] WEI Bo-cheng and LU Guo-bin. Introduction of Statistic Diagnostics [M], (in Chinese). Nanjing: Southeast University Press, 1991.

[15] WANG Tong-xiao and JIN Feng-xiang. Statistic Diagnostics and Data Processing of Gyrotheodolite Orientation [M], (in Chinese). Zuzhou: China Mining University Press, 1997.

(Edited by HE Xue-feng)