RESEARCH ON THE REGRESSION ALGORITHM ANALYSIS AND ITS APPLICATION

Shibing Sun[1, 2]* and Huan Zhao[3]

[1] School of Computer and Communication, Hunan University, China
*[*2]Software Department, Changsha Social Work College, 22 Xiang Zhang Road, Changsha Hunan, China*
Email: sunpine1979@yahoo.com.cn
[3]School of Computer and Communication, Hunan University, China
Email: hzhao@hnu.cn

ABSTRACT

An efficient method is proposed to diagnose a type of abnormal data. We first start with analyzing an example, carry through with development of the theory once more, and finally list the method steps and its application fields. Experiments show that we need more important and better ways to diagnose abnormal data and eliminate them along with the development of information technology and control technology. The quality of measured data is improved by the use of this technique.

Keywords: Abnormal data, Partial Least-Square (PLS) algorithm, Process error, Data quality

1 INTRODUCTION

The concept of steady estimating (or robust estimators) has a long history. As early as 1960, Tukey emphasized the importance of the steady estimating method: "It is an obvious expectation that could not bring on any serious consequences for neglecting the deviation of the ideal model, but under the strict model condition, the most superior statistical method also could be most superior under the approximate model. Unfortunately this hope is often extremely wrong; even some slight deviations also can have a more tremendous influence compared to our expectation."（Browne, M.W., & Bansal, P.K. 1998) We can regard the production process as a kind of information flow from the viewpoint of information theory and cybernetics. It concomitantly has the technology of batch information processing that reflects the state of the art and equipment advances, the mutual function and connection of various links, and conceals the regularity of optimal production. Batch data are manifested in this information that is the basis of the process control and optimization. However, the deficient erroneous-data are frequently derived from the process of data collection and flow, such as the malfunctions of a sensor, a switch, or recording instrument, the stochastic nature of the manual recording or bad data input, and so on.

No matter which kind of data we deal with, the data must reflect the reality as far as possible and must be accurate, reliable, and unabridged. Suppose we deal with some false data as true data: not only will the processing result be insignificant, but also it likely will result in poor decision-making and incorrect control based on the false information. Therefore, data must be distinguished principally into true data and false data regardless of its processing. The process of distinguishing among data is the error diagnosis and the result data are revised. In view of the fact that this domain is receiving more and more attention from theorists and actual data producers, this article analyzes the detection of abnormal data, proposes a method of the error data

diagnosis (The Steady Return Algorithm), and then introduces its application.

2 THE INTERFERENCE OF ERRONEOUS DATA IN REGRESSION ANALYSIS

Regression analysis is a popular and effective modeling method. Users usually draw a dimensional-dispersion pattern to distinguish error values in the cell datum process and then observe under different conditions. This method, however, does not work with multi-dimensional data. It has been maintained that the absolute value between the forecast value and the actual value (i.e. error) in regression analysis can be distinguished, but this conclusion may be unreasonable. Consider the examples in Table 1.

Suppose y and x have a linear relationship. Using the partial least-square return algorithm results in straight line L_1:

y=0.06833—0.08146x （See Figure 1）

order	x	y	L1		L2	
			y	r	y	r
1	—4	2.48	0.39	2.09	2.04	0.44
2	—3	0.73	0.31	0.42	1.06	—0.33
3	—2	—0.04	0.23	—0.27	0.08	—0.12
4	—1	—1.44	0.15	—1.59	—0.9	—0.54
5	0	—1.32	0.07	—1.39	—1.87	0.55
6	10	0	—0.75	0.75	—11.64	(11.64)
The surplus standard dispersion			σ_1 =1.55		σ_2 =0.55	
$\lvert r_{max} \rvert / \sigma$			1.35		1.00	

Table 1. Data sheet: y=0.06833—0.08146x

For each return value y_i and residual error r_i row in Table 1, the accurate return surplus dispersion is $|r_{max}|$=2.09. Because the largest absolute residual error value is σ_1=1.55, this result does not surpass the surplus standard deviation by more than two fold. Therefore, the result may be regarded as true data without an identified error according to the usual conventions. However, if we carefully observe Figure 1 again, we can discover in another straight line by excluding the 6th datum. If it is removed, we match a straight line with the other 5 data points and then obtain the line L_2: y= -1.87333-0.97767x (Figure 1).

Line L_2 and line L_1 are far from each other; moreover L_2 and the overwhelming majority of data points are all very close to each other; but the distance between L_1 and the majority of points is larger. Thus, we can draw the conclusion that the partial least-square return algorithm has two weaknesses:

First, an individual point may have a significant influence on the result in this data. Participation of the datum in the return obviously influences results, in this example, above the 6th data point.

Second, it is not possible to discover the error datum with the size of the residual error data.

The unusual 6th point residual error in the above example is quite small (0.75), but the biggest residual error (datum 1, r_i=2.09) is not the error datum. If an unusual point is suspected because of a large residual error or rejection, we will obtain a worse return result compared to L_1. Therefore, it is not reliable to judge an unusual value by the size of the residual error in the PLS algorithm; such an approach can go astray.

These problems exist mainly because that PLS algorithm minimizes the sum of the residual error. It regards all points equally and demands that the straight line approach each point to an extreme. However, when the data has an exceptional datum, it cannot obtain the correct regression equation because the PLS algorithm treats the unusual datum equally without discrimination. Therefore, when using this kind of the regression equation, the residual error is naturally unreliable. In addition, the observed value includes two parts, the independent variable x and the dependent variable y, but the residual error is the difference between the component of y and the returned value, which does not fully reflect the factor in the component of x. In the next section we will analyze this theoretically.

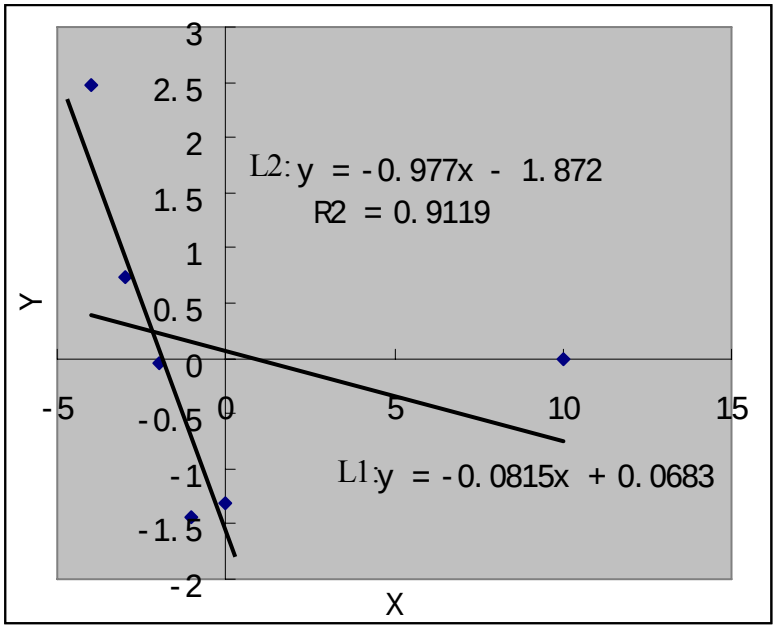

Figure 1. L_1: y=0.06833—0.08146x and L_2: y=-1.87333—0.97767x

3 THEORETICAL ANALYSIS OF THE ERROR PROCESS

Suppose the linear model $y_i = x_i'\beta + \varepsilon_i$ （i=1,2,……n), where x_i' is the transpose of x, (x_i, y_i), are the observed data, $x_i \in R_m$, $y_i \in R'$, $\beta = (\beta_1, \beta_2, ..., \beta_m)'$ are the regression coefficient to be estimated, and ε_i is the random error, namely $y = (y_1, y_2, ..., y_n)'$, $x = (x_1, x_2, ..., x_n)' = (x_{ij})_{n\times m}$. Therefore, the result returned by the PLS algorithm is

$$\hat{\beta} = (x'x)^{-1}x'y \qquad (1)$$

$$\hat{y} = x\hat{\beta} = x(x'x)^{-1}x'y = Hy \qquad (2)$$

where, $H = x(x'x)^{-1}x' = |\, x_i'(x'x)^{-1}x_j \,|_{n\times n} = (h_{ij})_{n\times n}$ （3）

The projection matrix of the x row vector in the line temper spatial, also called the hat matrix, the L^{th} residual error is:

$$r_l = y_l - y_l = y_l - \sum_{j=1}^{n} h_{ij} y_i = (1-h_{ij})y_l - \sum_{j\pm l} h_{ij} y_j$$

If y_i has an unusual value in the observed value of y, supposing its normal value is y_i^*,

$y_i = y_i^* + \Delta y$, the i^{th} residual datum is the following without any errors:

$r_i^* = (1-h_{ii})y_i^* - \sum_{j\neq i} h_{ij} y_j$ $[\, r_i = (1-h_{ii})(y_i^* + \Delta y_i) - \sum_{j\neq i} h_{ij} y_j \,]$

Because the error compels the i^{th} residual error to change, therefore $r_i - r_i^* = (1-h_{ii})\Delta y_i$

Therefore, although the error Δy_i is frequently larger, if h_{ii} is close to 1 (h_{ii} are the diagonal elements of the matrix in the projection matrix, therefore, $0 \leqslant h_{ii} \leqslant 1$), the difference between the corresponding r_i and r_i^* in the normal condition is not large.

Therefore the error in the i^{th} datum cannot be distinguished from the residual error r_i, and the change of the k(k≠i) residual error resulting from the error Δy_i is: $r_k - r_k^* = (1-h_{ki})\Delta y_i$ (k≠i)

If h_{ki} is bigger, the error instead is reflected in the k^{th} residual error. This means that it is unreliable to judge the unusual value in the corresponding points from the value size of the residual error r because the change of the residual error depends not only on the component errors of y but also on those of x, (h_{ij} is completely defined by x).

As shown above, there is a close relationship between the i^{th} datum and the corresponding h_{ij}. When h_{ij} is larger, this influence is more distinct. h_{ij} is generally called the leverage point or the latent influence point. Usually, an h_{ij} below 0.2 is regarded as being better, and as much as possible it should be at least below 0.5. In the above example, h_{66}=0.936 is extremely close to 1; therefore the return influence is very large, and it strays far into the x data.

4 ERROR DIAGNOSIS ALGORITHM

If we have an n by m dimension data set $x = (x_1, x_2, ..., x_n)' = (x_{ij})_{n\times m}$ based on above discussion and analysis, we can carry out the following algorithm steps to diagnose the unusual data:

Step 1: Calculate $x'x$

Step 2: Compute $x'x$ inverse matrix $(x'x)^{-1}$

Step 3: Confirm the threshold value F（0<F<1）to eliminate unusual values

Step 4: Compute $a_i = x_i'(x'x)^{-1}x_i$ i=1,2,…,n

Step 5: If there is an identical parent component in the new data set Z, then we calculate $a = z'(x'x)^{-1}z$, distinguish Z in contrast to the unusual datum and the normal datum with the Step 4.

In order to enhance the precision of the operation, we can standardize the primitive data set X and then carry on.

5 POSSIBLE APPLICATIONS FOR THE APPROACH

First, the solution in section 4 can be used to identify and reject unusual data in order to enhance the reliability, usefulness, and validity of the data.

Second, it can be used in breakdown diagnosis. If the signal examination method is good, we can gather the normal and unusual data for some application (i.e. industry control), confirm the a_i sector corresponding with the breakdown through the use of the previous algorithm, and then store $(x'x)^{-1}$. Using an online examination data as Z, we calculate $a = z'(x'x)^{-1}z$, finally ascertaining the degree of the breakdown according to size, which can then implement online, real-time quantitative analysis and a breakdown alarm of equipment performance.

Third, it can be used for planning and statistical management (Hur, Lee, & Baek, 2006; Liu, Wang, Su, & Tao, 2003). The project management, statistical report forms, and decision-making in a factory all depend on their access to needed data and their quality. However, these all have random error data, which can cause the administrators to miss the real efficiency of a factory. They can obtain reliable data with the observed values from data diagnosis and adjustment in the current capacity and other parameters (Browne& Bansal, 1998; Carrasco, 1998; Yuan & Li, 2004).

Fourth, it can be used in process signal tracing. Process data can be analyzed online using diagnostic and adjustment technology, equipment and work status of an instrument, which can be tracked and analyzed. Mistakes and malfunctions can be identified.

Fifth, it can be used in process control and optimization (Laursen & Stanciulescu, 2006; Yuan & Li, 2004). The combination of a flow simulator, optimizing algorithm, and diagnosis adjustment software can provide a reliable

process optimization plan. We can automatically and in real-time deposit and withdraw process data with diagnostic adjustment software and calculate time averages and adjustments to obtain uniform process data. Also we can input the adjusted data into a flow simulator to unify the most recent economic data on the operation in order to simulate and optimize process operation parameters to achieve the most economic value. Optimized processing parameter values are then available for new set-points process control. This is the online feedback control system. Optimization may be off-line or online and may also form an online closed-loop system with the optimizer and the control system linked together.

6 CONCLUSION

This article proposes an effective solution for diagnosing abnormal or unusual data. It begins with an example of actual data analysis, presents the theoretical inferential reasoning, and finally lists the algorithm steps and application domains of this method. As we know, diagnosis and resolving of unusual data, which can effectively enhance the quality of data by means of data adjustment technology, has become more important and urgent with the development of information technology and the control technology.

7 ACKNOWLEDGEMENT

I appreciate my teachers and some friends. Without their self-giving help, my experiments and thesis would not be completed. Thanks to Prof. Zhao Huan, Prof. Wang Jin Fang, Dr. Gong Zhong Liang and Dr. Yu Xin.

8 REFERENCES

Browne, M.W., & Bansal, P.K. (1998) Steady-state model of centrifugal liquid chillers. *International Journal of Refrigeration*, Aug, pp.343-358.

Carrasco, J. (1998) Bounding steady-state availability models with group repair and phase type repair distributions. *Proceedings of the 1998 3rd IEEE International Performance and Dependability Symposium (IPDS'98)*(pp. 193-214), Chicago, USA.

Charbonnier, S., Becq, G., & Biot, L.. (2004) On-Line Segmentation Algorithm for Continuously Monitored Data in Intensive Care Units. *IEEE Transactions on Biomedical Engineering*, March, pp.484-492.

Hur, J., Lee, H., & Baek, J..(2006) An intelligent manufacturing process diagnosis system using hybrid data mining. *Lecture Notes in Computer Science*, pp.561-575.

Laursen, T., & Stanciulescu I. (2006) An algorithm for incorporation of frictional sliding conditions within a steady state rolling framework. *Communications in Numerical Methods in Engineering*, April, pp.301-318.

Liu, F., Wang , X., Su, X., & Tao, W. (2003) Detection and reconciliation on the abnormal operation data based on redundancy measurement in a power plant. *China electrical engineering*, July, pp.204-207.

Yuan, Q. & Li, P. (2004) Optimal and robust design of unstable valve. *Proceedings of the 2004 American Control Conference (AAC)*. (pp. 4449-4454). New York, USA.