OUTLIER DECTION IN FUZZY LINEAR REGRESSION ABOUT PREDICYING GDP GROWTH

Xu Jialu & Lu Qiujun

Faculty of Science，University of Shanghai for Science and Technology，Shanghai 200093，China.

Email:ahxjialu@foxmail.com

ABSTRACT

The accuracy of regression model may have some warp when the observation data set exists outlier. We study to lessen the outlier effect in predicting GDP growth on the basis of fuzzy linear regression model in this paper. The result is eff ectively verified the performance of the model.

Keywords: *GDP; Fuzzy linear regression; Outlier; Linear programming.*

1. INTRODUCTION

On the basis of Zadeh's fuzzy theory [1], Tanaka and coworkers [2] were the first to put forward the concept of fuzzy linear regression. Since then, many approaches have been proposed in the literature of fuzzy regression. However, the Tanaka approach has a disadvantage which is extremely sensitive to the presence of outliers. The outlier is often difficult to avoid in practical application, they may occur because of specific reasons, such as abnormal time, the disturbance of external factor. Furthermore, gross error during the collection or recording observations is also one of the major reason. The number of outliers in data set may often more than 1, and unpredictable. Thus, it's significant to omit or lessen the effect of outlier on fuzzy linear regression. In regression analysis, the outlier problem can be dealt with two different, but not mutually exclusive, points of view: outlier detection [3-8] and robust estimation [9-15].

Literatures [16-17] often lack of the consideration of the fuzzy uncertainty about the GDP data or system, thus, the purpose of this paper is to study GDP growth based on fuzzy linear regression model [8]. The remainder of this paper is organized as follows. We introduce the fuzzy linear regression model and outlier detection in Section 2. In Section 3, GDP growth is analyzed with the method and the result is compared with some other existing fuzzy regression methods. The paper is then ended with conclusion.

2. FUZZY lINEAR REGRESSION MODEL AND OUTLIER DECTION

2.1 Fuzzy linear regression model [2]

In fuzzy linear regression, regression function expresses as follows:

$$\tilde{Y}_i = A^T x_i \qquad (1)$$

where $x_i = [1, x_{i1}, \cdots, x_{ip}]^T$ is a vector of input variables in the *i*th data, $x_{ij} \in R$, $i = 1,2,\cdots,n$, $j = 0,1,\cdots,p$;

$A = [A_0, A_1, \cdots, A_p]^T$ is a vector of symmetric triangular fuzzy parameters, $\tilde{A}_j$ represented by $\tilde{A}_j = (\alpha_j, c_j)_L$, with α_j and c_j are its central value and spread value, respectively. *L* represents the reference function, and $L(x) = \max(0, 1-|x|)$.

Thus, the fuzzy output $\tilde{Y}_i$ can also be described as $\tilde{Y}_i = (\sum_{j=0}^{p} a_j x_{ij}, \sum_{j=0}^{p} c_j |x_{ij}|)_L$, $i = 1,2,\cdots,n$, $j = 0,1,\cdots,p$. The membership function for $\tilde{Y}_i$ can be represented by the following symmetric triangular function**:**

$$\mu_{\tilde{Y}_i}(y_i)=\begin{cases}0, & x=0, y=0\\ 1-\dfrac{\left|y_i-x_i^T\alpha\right|}{c^T\left|x_i\right|}, & x\neq 0\\ 1, & otherwise\end{cases} \qquad (2)$$

where $c=[c_0,c_1,\cdots,c_p]^T$, $\alpha=(\alpha_0,\alpha_1,\cdots,\alpha_p)^T$.

2.2.Outlier detection and adjustment in fuzzy linear regression model [8]

Above all, apply the ordinary linear regression on the entire original data. From the empirical rule, it will be assessed as a proper curve when the coefficient of determination R^2 of fitting curve is more than or equal 0.8. Otherwise, from first to end in turn, put away one observation and fit a curve by ordinary regression to the other remaining data, accompany with keeping the corresponding R-square amount in each phase, delete the observation by ignorance of which the maximum of R-square is achieved. The process will over if R-squared statistic satisfies the empirical rule. At last, We achive $\hat{y}_i$ in the final ordinary regression equation, $i=1,2,\cdots,n$.

Fuzzifing y_i, 3σ principle is introduced, concern the symmetric triangular fuzzy number $(\hat{y}_i,3\sigma)_L$, where $\sigma=\sqrt{Var(y_i)}$. $Var(y_i)$ denotes the variance of y_i. μ_i is referred to the membership of y_i to $(\hat{y}_i,3\sigma)_L$. Then we adjust y_i as follows:

$$\begin{aligned}&\text{If } \mu_i=0, \text{then delete } y_i;\\ &\text{If } \mu_i=0, \text{then } y_i^*=\hat{y}_i+\frac{(y_i-\hat{y}_i)(1-\mu_i)}{2}.\end{aligned} \qquad (3)$$

Thus, by substituting the new data y_i^* instead of the original y_i, we obtain a new data set, n_1 denotes the number of new data set.

After achieve the new data set, the parameters can be derived by the following linear programming:

$$\min\sum_{i=1}^{n_1}\left[\left|\sum_{j=0}^{p}\left(a_jx_{ij}+(1-h)c_j\left|x_{ij}\right|\right)-y_i^*\right|+\left|\sum_{j=0}^{p}\left(a_jx_{ij}-(1-h)c_j\left|x_{ij}\right|\right)-y_i^*\right|\right]$$

$$s.t.\ \sum_{j=0}^{p}a_jx_{ij}+(1-h)\sum_{j=0}^{p}c_j\left|x_{ij}\right|\geq y_i^*, i=1,2,\cdots,n_1$$

$$\sum_{j=0}^{p}a_jx_{ij}-(1-h)\sum_{j=0}^{p}c_j\left|x_{ij}\right|\leq y_i^*, i=1,2,\cdots,n_1 \qquad (4)$$

$$\sum_{j=0}^{p}c_j\left|x_{ij}\right|\geq 0, i=1,2,\cdots,n_1$$

$$a_j\in R,\ j=0,1,\cdots,p$$

where, $n_1\leq n$. $0\leq h\leq 1$, the h-level set is the threshold level to be chosen flexibly.

In order to make a comprehensive comparison, we introduce a metric for evaluating the accuracy of a method in predicting the fuzzy response as

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|UB_i - LB_i\right| \tag{5}$$

where, *UB* and *LB* denote the upper and lower band estimate of fuzzy dependent variable , respectively. If *MAD* is smaller, outlier may have less effect on the regression model.

3. EMPIRICAL RESEARCH

In the whole economy system, the gross domestic product (GDP) data is effected by many factors, we consider production function to analysis GDP in this paper. We select GDP growth (denoted by *Y*) as the dependent variable, Then the explanatory variables are total investment in fixed assets growth (denoted by X_1), employment growth (denoted by X_2), electricity consumption growth (denoted by X_3) and national finance funding of science and technology growth(denoted by X_4).We utilize the data of China in the period of 1996through 2011 [18], the data of explanatory variables are processed because of no direct data (see Table 1).

Table 1 the data of GDP growth and explanatory variables

NO.	Year	*Y*(%)	X_1(%)	X_2(%)	X_3(%)	X_4(%)
1	1996	10.01	14.46	1.30	7.39	15.28
2	1997	9.30	8.85	1.26	4.83	17.30
3	1998	7.83	13.89	1.17	2.78	7.26
4	1999	7.62	5.10	1.07	6.09	24.01
5	2000	8.43	10.26	0.97	9.48	5.83
6	2001	8.30	13.05	0.99	8.63	22.19
7	2002	9.08	16.89	0.66	11.60	16.05
8	2003	10.03	27.74	0.62	16.53	15.73
9	2004	10.09	26.83	0.72	15.45	15.95
10	2005	11.31	25.96	0.52	13.51	21.88
11	2006	12.68	23.91	0.44	14.63	26.49
12	2007	14.16	24.84	0.46	14.42	26.49
13	2008	9.63	25.85	0.32	5.59	22.25
14	2009	9.21	29.95	0.35	7.21	25.50
15	2010	10.45	12.06	0.37	13.24	28.07
16	2011	9.30	23.76	0.41	12.08	14.30

Above all, the ordinary regression can be executed by original data as:

$$\tilde{Y} = 3.294 + 0.076X_1 + 1.259X_2 + 0.197X_3 + 0.114X_4 ;\text{with } R^2 = 0.60 . \tag{6}$$

where,the R-squared is less than 0.8. Then in terms of the R-squared statistic, we put away the fourteenth, eighth, ninth and sixth data in fourth loop in turn. Then we obtain the final ordinary regression as below:

$$\tilde{Y} = -0.740 + 0.179X_1 + 3.376X_2 + 0.271X_3 + 0.134X_4 ;\text{with } R^2 = 0.84 . \tag{7}$$

$\hat{y}_i$ can be calculated by equation (7), then we obtain y_i^* (see Table 2)by adjusting $\hat{y}_i$.

Table 2 the value of y_i^*

NO.	1	2	3	4	5	6	7	8
y_i^*	10.29	8.76	7.44	8.56	7.77	9.89	9.76	12.11
NO.	9	10	11	12	13	14	15	16
y_i^*	12.10	12.18	12.55	12.93	9.48	10.81	10.05	10.03

At last, we obtain the fuzzy regression equation by implement the equation (4):

$$\tilde{Y} = (0.1094, -4.3844) + (0.1637, 0.0878)X_1 + (2.9034, 2.3414)X_2 + (0.2575, 0.1042)X_3 + (0.1241, 0.0384)X_4, h = 0 \quad (8)$$

We implement the equation (4) on the original data , the result is given as:

$$\tilde{Y} = (1.5745, -16.3690) + (0.1511, 0.2795)X_1 + (2.5396, 8.0402)X_2 + (0.2464, 0.4934)X_3 + (0.0688, 0.2353)X_4, h = 0 \quad (9)$$

Compare with 2000 and 2002, there is no obvious change in the total investment in fixed assets growth, employment growth and electricity consumption growth in 2001, and national finance funding of science and technology growth in 2001 is considerably larger 2000 and 2002, however, the GDP growth is smaller than in 2000 and 2002. Thus, the data in 2000 may be underestimated. Furthermore, China was seriously affected by the Severe Acute Respiratory Syndrome (SARS) and global financial crisis in 2003 and 2009 respectively, datas of 2003 and 2009 may be outliers in consideration of the huge external factors.Therefore, put away the fourteenth, eighth, ninth and sixth data in the ordinary regression is reliable.

Compare the two approaches with a similar index *MAD* . The *MAD* criterion for the model (8) and (9) are respectively 1.55 and 8.59. Obviously, the index is reduced about five times by adjusting the outlier. Moreover, as the deviation between fitted center values and observed values are less than twenty percent of observed values, the model is acceptable.

4. CONCLUSION

As can be seen from the fuzzy model, employment growth has the most significant contribution to GDP growth, along with electricity consumption growth, total investment in fixed assets growth, and national finance funding of science and technology growth. It shows that China is still a labor-intensive society, and the development of economic crucial depend on labor-intensive industry. It calls for change in the pattern of economic development to further enhance the economic competitiveness.

In this paper, we apply ordinary regression along with 3σ princple to lessen the effect of outlier on fuzzy linear regression of GDP growth. Furthermore, the prediction of GDP is not a crisp value, but a range that is present between the upper bond and the lower bond, which overcomes the problem caused by inaccuracy of crisp values.

5. ACKNOWLEDGEMENT

This research was supported by the Doctoral Scientific Research Funds of University of Shanghai for Science and Technology (1000341001).

6. REFERENCES

[1] Zadeh L A . Fuzzy sets. Information and Control,1965,8(3):338-353.

[2] Tanaka H, Uejima S, Asai K.Linear regression analysis with fuzzy model. IEEE Transactions on Systems ,Man ,and Cybernetics,1982,12(6):903-907.

[3] Hung W L, Yang M S. An omission approach for detecting outliers in fuzzy regression models.Fuzzy Sets and Systems ,2006, 157(23):3109-3122.

[4] Coppi R, D'Urso P, Giordani P, et al. Least squares estimation of a linear regression model with LR fuzzy

response. Computational Statistics and Data Analysis, 2006,51(1):267-286.

[5] Peters G. Fuzzy linear regression with fuzzy intervals. Fuzzy Sets and Systems ,1994, 63 (1):45-55.

[6] Chen Y S. Outliers detection and confidence interval modification in fuzzy regression. Fuzzy Sets and Systems,2001,119(2):259-272.

[7] Gladysz B, Kuchta D. Outliers detection in selected fuzzy regression models //Masulli F,Mitra S, Pasi G. Applications of Fuzzy Sets Theory: 7th International Workshop on Fuzzy Logic and Applications,WILF 2007, Camogli, Italy, July 2007,Proceedings. Berlin Heidelberg: Springer,2007:211-218.

[8] H Shakouri,R Nadim.Outlier detection in fuzzy linear regression with crisp input–output by linguistic variable view.Applied Soft Computing,2013, 13(1): 734–742.

[9] Watada J, Yabuuchi Y. Fuzzy robust regression analysis based on a hyperelliptic function // IEEE Neural Networks Council.Proceedings of 1995 IEEE International Conference on Fuzzy Systems:The International Joint Conference of the 4th IEEE International Conference on Fuzzy Systems and the 2nd International Fuzzy Engineering Symposium: March 20-24, 1995, Yokohama, Japan. New York:IEEE, 1995:1841 -1848.

[10] Özelkan E C ,Duckstein L.Multi-objective fuzzy regression :a general framework.Computers and Operations Research, Special Issue on Artificial Intelligence and Decision Support with Multiple Criteria ,2000,27(7):635-652.

[11] Torabi H, Behboodian J. Fuzzy least-absolutes estimates in linear models. Communications in Statistics:Theory and Methods,2007, 36(10):1935-1944.

[12] Modarres M, Nasrabadi E, Nasrabadi M M . Fuzzy linear regression analysis from the point of view risk. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2004,12(5):635-649.

[13] Modarres M, Nasrabadi E, Nasrabadi M M . Fuzzy linear regression models with least square errors. Applied Mathematics and Computation,2005,163(2):977-989.

[14] Nasrabadi E, Hashemi S M. Robust fuzzy regression analysis using neural networks. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2008,16(4): 579 -598.

[15] Hu Y C .Functional-link nets with genetic-algorithm-based learning for robust nonlinear interval regression analysis. Neurocomputing, 2009, 72(7-9):1808-1816.

[16] Meng L ,W X L. An Estimation of the Reliability of Statistic Data on China‘s Economic Growth. Economic Research Journal, 2000 (10) : 3-13.

[17] Cai Fang and LuYang, 2013. Population Change and Resulting Slowdown in Potential GDP Growth in China, 21(2), 1-14.

[18] National Bureau of Statistics of the People's Republic of China. China Statistical Yearbook. 2011. Beijing: China Statistics Press, 2012.