Combining Numeric Method and Visualization Method Together to Analyze Big Data and the Prediction of the Rate of Accidental Death In China's Coal Mining Industry

Abstract:

In this paper, we want to introduce an enhanced least square method. We will utilize this method to analyze the give data, which is the number of deaths annual in colliery accidents in China since 2005. And we will predict the future performance, offering an opinion about the current measures for safety precautions in coal industry.Analyzing the rules of the big data can not only help analyze the situation, but predict the trends, allowing an improvement of probability in decision making.In this research, we will use the Standard Total Deviation and Pearson Correlation Coefficient analysis methods to conduct the error analysis.

Keywords: technological development, big data, inherent laws, nonlinear equation

1 Introduction

The boom in technological development such as network has fueled a massive amount of data from heterogeneous sources, which is still significantly increasing every day [1].

Though most of the data collected are not organized and seems to be irregular, considering that data contains so much hidden information, the significance of analyzing these data should be no doubt. For instance, by collecting and analyzing the data, scientists are able to discriminate the signals from the noise [2]. And for economy, advanced techniques and methods are also required nowadays. By studying the data, econometricians are able to extract the inherent laws, understand the current situation and even help make wiser decisions[3].

Analyzing the rules of the big data can not only help analyze the situation, but predict the trends, allowing an improvement of probability in decision making. The prediction could quickly reflect that if the current decision or policy is efficient and useful or not.

Adjustment can be made timely by analyzing the prediction [4].
Under the trend in scientific and economic development, for Computer Scientists and Mathematicians, advancing the data analysis methods for prediction gives significance to the society [5]. For over two hundred years, scientists had tried to find out as well as enhance the data analysis methods. One of the methods is the leaner squares method. The Least Square Method is one of the classic data analysis methods which was first introduced in 1805 by Legendre and also in 1809 by Gauss. It was utilized for various kinds of data analysis and enhanced until nowadays [6][7].
In this paper, we want to introduce an enhanced least square method. We will utilize this method to analyze the give data, which is the number of deaths annual in colliery accidents in China since 2005. And we will predict the future performance, offering an opinion about the current measures for safety precautions in coal industry.
The paper has five sections. Section 2 presents the method. Section 3 presents the case study. Since the data cannot be expressed as linear equation, a conversion between linear equation and nonlinear equation is required. And then we will use the algebraic method to find out the linear equation, and then convert it back into the actual nonlinear equation. We will predict the future data. Section 4 presents the error analysis and discussion. We will use the error analysis methods to see how this method can perform the prediction close to the actual data.
If the method works well, then the prediction can provide a suggestion for the efficiency of the precautionary measures in colliery accidents. The final section presents the summary and conclusion.

2 Numeric Analytic Algorithms

In this paper, we will use the enhanced Least Square Method. Detailed introduction is presented as follow:

Table 1. Formulas

M_{11}	$\sum_{i=1}^{n} x_i^2$
M_{12}	$\sum_{i=1}^{n} x_i$
M_{13}	$\sum_{i=1}^{n} y_i x_i$

M_{21}	$\sum_{i=1}^{n} x_i$
M_{22}	$\sum_{i=1}^{n} 1$
M_{23}	$\sum_{i=1}^{n} y_i$
L	$(M_{11}*M_{22}) - (M_{12}*M_{21})$
L_1	$(M_{13}*M_{22}) - (M_{12}*M_{23})$
L_2	$(M_{11}*M_{23}) - (M_{21}*M_{13})$
A	L_1/L
B	L_2/L

3 Case Study

We will utilize this method to analyze a case. The procedures will present how this method works in detail.

3.1 Data

The total number of deaths caused by colliery accidents in China from 2005 to 2018 are listed in Table 1. We will analyze this 14-year-data and predict the future five year's number of deaths annual.

Table 2. Total number of deaths caused by colliery accidents annual

Year	x_i	Number of deaths (y_i)
2005	1	5938
2006	2	4746
2007	3	3786
2008	4	3215
2009	5	2631
2010	6	2433
2011	7	1973
2012	8	1384
2013	9	1067
2014	10	931

2015	11	588
2016	12	538
2017	13	375
2018	14	221

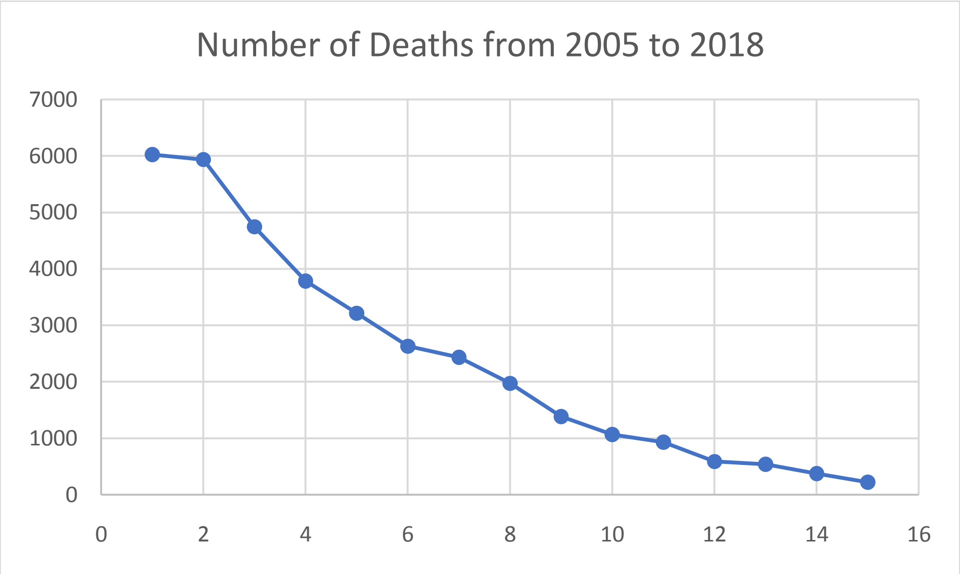

Fig. 1. Number of Deaths from 2005 to 2018

3.2 Given Formulas

Since the data is cannot be expressed in a linear equation, we will use a mathematical transformation to turn the nonlinear equation into a linear equation, and we will turn the solved linear equation back to a nonlinear equation in the end.

From figure 1, we can see that it is a nonlinear exponential function. We assume that the function to express the data should be as follow:

$\mathrm{Y}(x;\ \mathrm{a},\ b) = \mathrm{a} * e^{bx}$

We then take a logarithm operation to both sides of equation and get:

$\mathrm{G}(x;\ \mathrm{b},\ \mathrm{c}) = bx + c$

$\mathrm{G}(x;\ \mathrm{b},\ \mathrm{c}) = \ln \mathrm{Y}(x;\ \mathrm{a}, b)$, where $c = lna$

3.3 Convert the Data

We then convert the given data (y_i) to $\ln(y_i)$. The converted data are listed in Table 2. And the visualization of the converted data is shown in Figure 2. We can see that figure 2 presents a linear function pattern.

Table 3. Converted data from original data

Year	x_i	Number of deaths $\ln(y_i)$
2005	1	8.689127655
2006	2	8.465057437
2007	3	8.239065332
2008	4	8.075582637
2009	5	7.875119281
2010	6	7.796880343
2011	7	7.587310506
2012	8	7.232733136
2013	9	6.972606251
2014	10	6.836259277
2015	11	6.376726948
2016	12	6.28785856
2017	13	5.926926026
2018	14	5.398162702

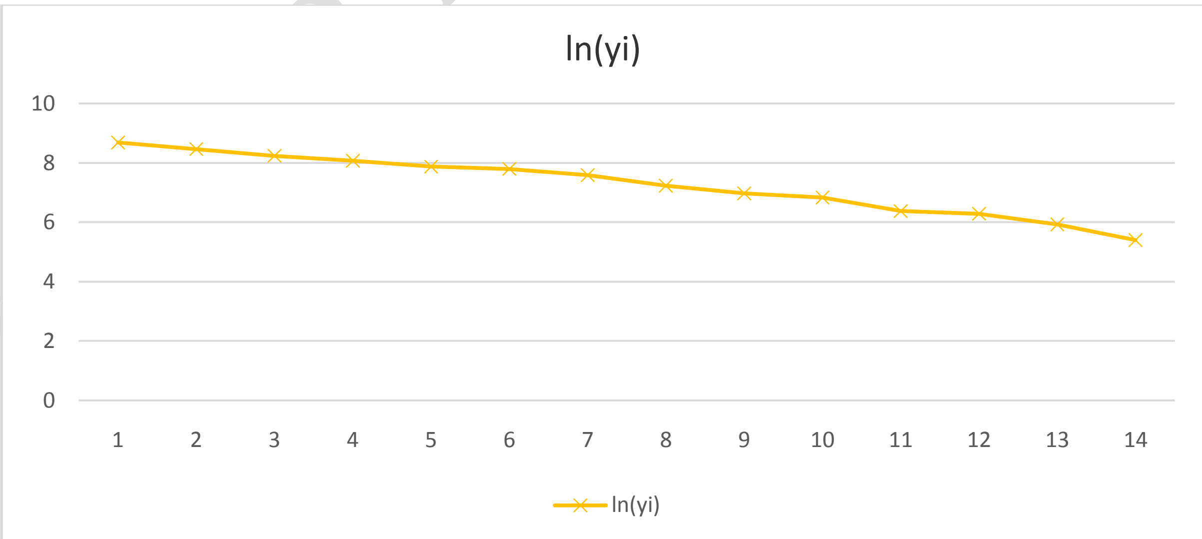

Fig. 2. Converted data, $\ln(y_i)$, from 2005 to 2018

3.4 Compute the Linear Function

We use a computer program (in this case, it is Java) to compute the variables and convert the data. Table 4 lists the output values and Table 5 lists the converted data.

Table 4. Output Values

M11	1015
M12	105
M13	709.10736
M21	105
M22	14
M23	101.75941
L	3185
L1	-757.2344
L2	28829.523
A	-0.2377502
B	9.051656

Table 5. Converted Original Data $\ln(y_i)$

Year	x_i	$G(y_i)$
2005	1	8.813905715942383
2006	2	8.576155662536621
2007	3	8.33840560913086
2008	4	8.100654602050781
2009	5	7.8629045486450195
2010	6	7.625154495239258
2011	7	7.387404441833496
2012	8	7.149654388427734
2013	9	6.9119038581848145
2014	10	6.674153804779053
2015	11	6.436403274536133

2016	12	6.198653221130371
2017	13	5.960903167724609
2018	14	5.723153114318848

3.5 Prediction Based on the Linear Function

We then predict the future values in five years using the function we computed ($\ln(y_i)$ = Ax+B). The predict data are listed in Table 6.

Table 6. Predict data using $\ln(y_i)$=Ax+B

Year	x_i	$\ln(y_i)$
2019	15	5.485403060913086
2020	16	5.247652530670166
2021	17	5.009902477264404
2022	18	4.772151947021484
2023	19	4.534401893615723

3.6 Comparison of Converted Data y_i and Predicted Data G(x_i)

We use the Microsoft Excel to visualize the original data and the theoretically data in order to see if the two lines are close or not. The patterns are shown in Figure 3, and we can see that the two lines match each other.

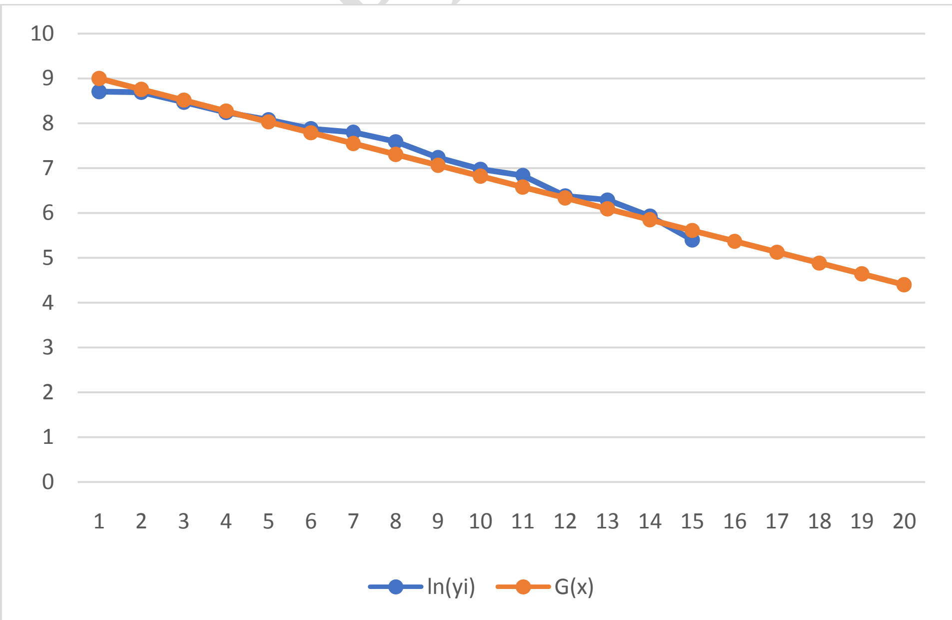

Fig. 3. Comparison of original converted data G(y_i) and theoretically data

3.7 Transform the Linear Function back to Nonlinear Function

Since

$G(x; b, c) = bx + c$

$G(x; b, c) = \ln Y(x; a, b)$, where $c = \ln a$

We have

$a = e^c = 9.051656$

$Y(x; a, b) = e^{G(x; b, c)} = 9.051656 * e^{-0.2377502x}$

3.8 Prediction Based on the Nonlinear Function

Now we have the nonlinear function to predict the future five years value. The converted are listed in Table 7, while the predictive data are listed in Table 8. We also use the Microsoft Excel to draw the comparison patterns of the original data and predictive data are present in Figure 4. From Figure 4, we can see that most of the predictive values are close to the original values.

Table 7. Converted Original Data to y_i

Year	x_i	$Y(y_i)$
2005	1	6727.141328389617
2006	2	5303.675578285504
2007	3	4181.415615974132
2008	4	3296.6265258076223
2009	5	2599.0591342177804
2010	6	2049.097016300431
2011	7	1615.50733012115
2012	8	1273.665381831056
2013	9	1004.1571964697285
2014	10	791.6772568454153
2015	11	624.1579822165451
2016	12	492.0858637735224
2017	13	387.9603593284341
2018	14	305.867767407298

Table 8. Predicted data

Year	x_i	$Y(y_i)$
2019	15	241.14606349459257

2020	16	190.11944464590994
2021	17	149.8901177419238
2022	18	118.17327115283388
2023	19	93.1677743959138

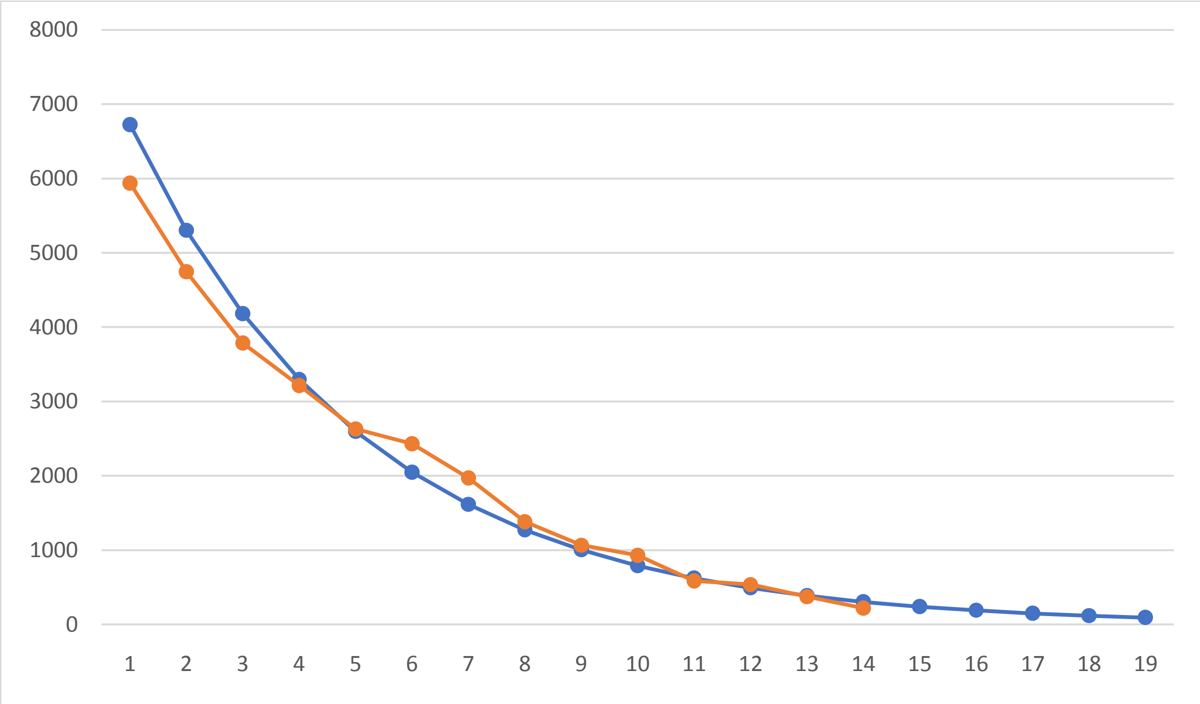

Fig. 4. Comparison of original data $y_i(red)$ and the Predicted data *Y (xi), (blue)*

4 Error Analysis and Discussion

Though we see that the two lines (actual data curve and the predictive data curve) are close in most of the time, we still need to verify how accurate the prediction is, using some error analysis methods.

In this research, we will use the Standard Total Deviation and Pearson Correlation Coefficient analysis methods to conduct the error analysis. Analysis formulas and results are listed in Table 9.

Table 9. Statistical analysis of errors and Pearson correlation coefficient

Error type	Error formula	Error value
Total difference of original and theoretically data	$\sum_{1}^{14}(y_i - Y(x_i))$	-826.094337
Total least square error	Q(X) = $\sum_{1}^{14}(y_i - Y(x_i))^2$	1419285.004
Mean value of original data	$\bar{y} = \frac{1}{14}\sum_{1}^{14} y_i$	2130.428571
Total difference of original data and mean value	$\sum_{1}^{14}(y_i - \bar{y})$	-5.45697E-12
Total variance	$\delta_{14}^2 = \frac{1}{14}\sum_{1}^{14}(y_i - \bar{y})^2$	40391957.43
Standard total deviation	δ_{14}	6355.466736
Pearson correlation coefficient	$\frac{\sum_{1}^{14}(x_i - \bar{x})(y_i - \overline{y})}{\sqrt{\sum_{1}^{14}(x_i - \bar{x})^2}\sqrt{\sum_{1}^{14}(y_i - \bar{y})^2}}$	0.993475881

5 Conclusion

We introduced an enhanced Least Square Method for analysis and prediction of Big Data. We used the number of deaths caused by colliery accident in China as data in our case study. In the case study, we show the readers how to operate the transformation between linear equation and nonlinear equation, and how to solve the functions using the data analysis method we introduced. In the end, we compare the predictive data and the original data, and conduct an error analysis for this data analysis method. Given the results of the analysis, we come to a conclusion that this method is able to analyze Big Data with a high accuracy.

However, this does mean we can stop trying to improve the methods for data analysis and predict. Challenges in the data analysis for Big Data should be considered significantly. Whenit comes to Big Data whose volume is huge, the complexity is much higher than the case we study in this paper. In this sense, we need to try to enhance the current method for a more complex and irregular data.

References

[1]Oussous, A., Benjelloun, F., Lahcen, A. A., &Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences, 30*(4), 431-448. doi:10.1016/j.jksuci.2017.06.001

[2]Wu, Z., & Huang, N. E. (2009). Ensemble Empirical Mode Decomposition: A Noise-Assisted Data Analysis Method. *Advances in Adaptive Data Analysis, 01*(01), 1-41. doi:10.1142/s1793536909000047

[3] Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives, 28*(2), 3-28. doi:10.1257/jep.28.2.3

[4] Elgendy, N., &Elragal, A. (2016). Big Data Analytics in Support of the Decision Making Process. *Procedia Computer Science, 100*, 1071-1084. doi:10.1016/j.procs.2016.09.251

[5]Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137-144. doi:10.1016/j.ijinfomgt.2014.10.007

[6]Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Smolley, S. P. (2017). Least Squares Generative Adversarial Networks. *2017 IEEE International Conference on Computer Vision (ICCV)*. doi:10.1109/iccv.2017.304

[7]Guo, Q., & Ye, P. (2018). Error analysis for lql^{q}-coefficient regularized moving least-square regression. *Journal of Inequalities and Applications, 2018*(1). doi:10.1186/s13660-018-1856-y

[8] Zou YJ, Chen Z, Xu J (2016) Binding Visualization Method and Numeric Method Together to Analyze Large Data – With a Case Study. British Journal of Mathematics and Computer Science 15: 1-10.

[9] Zou Y, Jun S Huang, Tong Xu, Anne Zou, Bruce He, et al (2019) Prediction of Coal Mining Accidental Death for 5 Years based on 14 year's Data Analysis. In the Book Series of Advances in Intelligent and Computing 1143: 281-289.

[10] Zou Y, Xin Luo, Anne Zou (2021) Big Data and Machine Learning: Algorithms for Analysis, with Case Studies. Proceedings of the 10th International Conference on Information Sciences, March 6 – 7, 2021, Tokyo, Japan. @2021 INTERNATIONAL Information Institute 23-28.

[11] Zou YJ (2017) A New Software Methodology for Decision-Making Based on Big Data and Machine Learning, Long Abstract of Invited Speech at 2017 IAENG International Conference on Computer Science, March 15 – 17, 2017, Hong Kong. Lecture Notes of Engineering and Computer Science 2017: Book I & II, pp lvii - lviii.