ISSN: 2754-6705

Journal of Mathematical & Computer Applications

SCIENTIFIC Research and Community

Review Article

Open Access

Prediction of Guam's Registered Voters Based on Data Analysis - A Case Study

Yousuo J Zou* and Jayvee Andrada

Computer Science Program, College of Natural and Applied Sciences, University of Guam, Guam 96923, USA

ABSTRACT

In various disciplines such as education, life science, meteorology, and environmental studies and as well as industries and businesses, there is a need to analyze large amounts of data (Big Data). Although, to find an effective and high-performance analytic algorithm that can predict future performance of selected data areas is still a challenge among Data Scientists, especially Computer Scientists and Mathematicians. In this paper, we will combine both a data visualization method and numeric analysis method in order to analyze large data (Big Data). This paper also includes a case study selected from the 2013 Guam Statistical Yearbook of Registered Voters as the original data, step by step, to show how to analyze a nonlinear data pattern and predict the future registered voters. Computational and mathematical efforts in this paper suggest that the proposed methods (or algorithms) are efficient in data analysis and data prediction using the nonlinear regression method.

***Corresponding author**

Yousuo J Zou, Computer Science Program, College of Natural and Applied Sciences, University of Guam, Guam 96923, USA.

Received: February 08, 2024; **Accepted:** February 13, 2024, **Published:** February 23, 2024

Keywords: Big Data Analysis, Visualization Methods, Nonlinear Regression, Case Study, Prediction, Guam, Registered Voters

Introduction

In the studies of data science include data collection, data storage, data transport / communication, data usage, data analysis, data visualization and knowledge discovery, etc. There are huge amounts of data (Big Data) that are generated every day in industries, businesses, education, science and engineering which need to be analyzed [1]. Although there are some large software systems that have been developed for massive data processing, such as traditional SPSS, SAS and newest MapReduce and Hadoop, etc. However, any large and powerful data processing software must use high-performance and efficient data analysis algorithms, which is still a challenging task for Data Scientists, especially Computer Scientists and Mathematicians, to find fast and efficient data analysis methods (or algorithms) that can quickly and accurately produce data analysis results for knowledge discovery or for prediction of future performance in the selected data area [2].

Data Scientists have been trying hard to apply the most advanced technologies into data analytics. One of the new technology that has been applied in data analysis is the digital visualization technology [3]. As the progresses in digital multimedia technology, Data Scientists now are able to use digital graphics, animation, audio and video techniques in data analysis [4]. They integrate the newest and oldest data analysis methods together, or bind the data visualization methods and traditional numeric methods together to analyze massive or big data [5,6].

The numeric data analysis methods, such as the Least Square Methods, were long been used for data analysis [5,6]. It started to be used in data analysis as early as in 1805 by A. M. Legendre and in 1809 by C. F. Gauss. There are many research articles and monographs in modern days and in history using advanced the numeric data analysis methods that made the Least Square Methods become a reliable and efficient data analysis tool [7-10].

In this paper, we want to share our research methods and experience on binding the visualization method and numeric analysis method together to analyze the selected registered Guam voting data. Included in this paper is a case study that shows, step by step, how our methods of data analysis work well with the data and how the methods can successfully make predictions for the future performances on Guam.

Case Study

In order to show readers how our data analysis methods work, we included a step by step case study to see how to use our analysis methods.

Data

Registered Voters are the given data in Table 1 [11], which are real world 18 year's data (1950 – 1968)

Table 1: Registered Voters on Guam

Year	x_i	y_i Registered Voters
1950	1	5415
1952	2	7988
1954	3	10093
1956	4	11987
1958	5	17077
1960	6	23483
1962	7	28854
1964	8	35207
1966	9	42664
1968	10	53065

Task

Using the data analysis results to predict the future 6-years (1969 – 1974) of increase

Data Analysis Procedure

Step 1

Since the given data is not too large, we will use Microsoft Excel as our visualization tool.

By writing a computer program to perform our numeric data analysis algorithms, we can see from Table 1 that our selected data visualization software tool shows the function pattern of the given data (Figure 1 in the following).

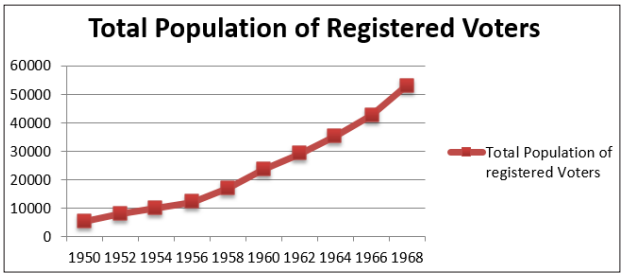

Figure 1: Registered Voters 1950–1968

Step 2

The data shows a functional pattern of a nonlinear exponential function. We will then use the following function pattern to describe the data

$$Y(x; a, b) = ae^{bx} \qquad (1)$$

Step 3

If we substitute the Equation (1) into the Least Square Equations, we will get two nonlinear algebraic equations with a and b as the variables. We will then use a mathematical transformation to turn Equation (1) into a linear equation. We take a logarithm operation to both sides of Equation (1), to get

$$G(x; b, c) = bx + c \qquad (2)$$

Where

$$G(x; b, c) = Ln\ Y(x; a, b) \quad c = Ln\ a \qquad (3)$$

Now we have successfully turned a nonlinear Equation (1) into a linear Equation (2).

Step 4

By using computer program (Java programming tool) we will convert $g(x_i) = ln(y_i)$ and list all the new converted data in the following Table 2.

Table 2: Converted data from original data y_i

Year	x_i	Ln(y_i)
1950	1	8.5969
1952	2	8.9857
1954	3	9.2196
1956	4	9.3916
1958	5	9.7455
1960	6	10.0640
1962	7	10.2700
1964	8	10.4690
1966	9	10.6611
1968	10	10.8793

Step 5

Now using the selected data visualization tool (Microsoft Excel) and the data in Table 2 to show the data pattern of the new dataset (Figure 2). It shows that this is a very good linear function pattern as described as in Equation (2).

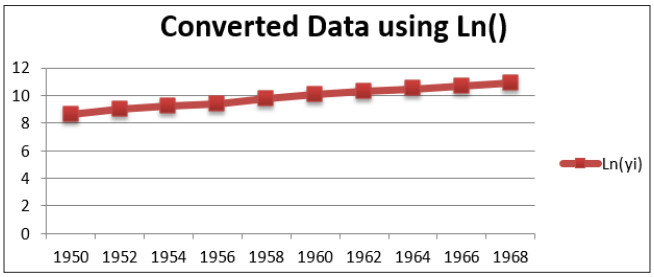

Figure 2: Function Pattern of Converted Data, Ln(y_i), in 1950 –1968

Step 6

Substitute Equation (2) into the Least Square Equations, we can get two linear algebraic equations, Equations (4) with different values of $M_{11}, M_{12}, M_{13}, M_{21}, M_{22}, M_{23}$ since the data in Table 2 are different from data in Table 1 [8].

$$\begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \begin{pmatrix} b \\ c \end{pmatrix} = \begin{pmatrix} M_{13} \\ M_{23} \end{pmatrix} \qquad (4)$$

Where $M_{11} = \sum_1^{10} x_i^2$ $M_{12} = M_{21} = \sum_1^{10} x_i$

$M_{22} = \sum_1^{10} 1$ $M_{13} = \sum_1^{10} y_i\ x_i$ $M_{23} = \sum_1^{10} y_i$

Using the above constants and the regular linear algebraic routine to solve Equation (4) and get the numeric values of b and c.

$b = 0.2513$

$c = 8.4459$

Therefore, we have solved new linear Equation (2):

$$G(x; b, c) = 0.2513\ x + 8.4459 \qquad (5)$$

Step 7

Using the selected data visualization tool (Microsoft Excel in this case) and the data analysis resulted function pattern, Equation (5), to draw a predictive linear line, and compare with the linear line based on the data in Table 2 to see the two lines match closely or not (Figure 3). Since the Least Square Methods are the optimal and the best fitting for the data being used, we can see from Figure 3 that the data line and the theoretical line are matched very well

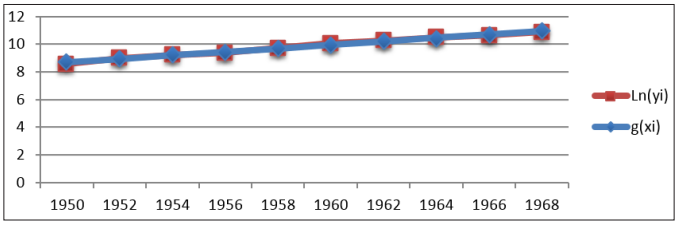

Figure 3: Comparison of Converted Original Data Ln(y_i) and Predictive Data G (x_i)

Step 8

Now we will turn the linear function pattern back into original nonlinear function pattern,

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

$$a = e^{c} = 4655.9442$$

So the original non-linear exponential function pattern will be

$$Y(x; a, b) = 4655.9442\, e^{0.2513x} \qquad (6)$$

Step 9

Using the predictive *Equation* (6) and Microsoft Excel to draw the exponential curve. We can now compare the curve generated by original data given in Table 1 (Please see Figure 4). From Figure 4, we can see the theoretical function pattern (red line) very closely matches the data curve (blue line). In the Figure 4, we also employ the Formula (6) above to make the prediction of the future performance in 1970, 1972, and 1974. Is this prediction accurate enough or not? We can use the actual data or performance of the Yangtze River Port to verify.

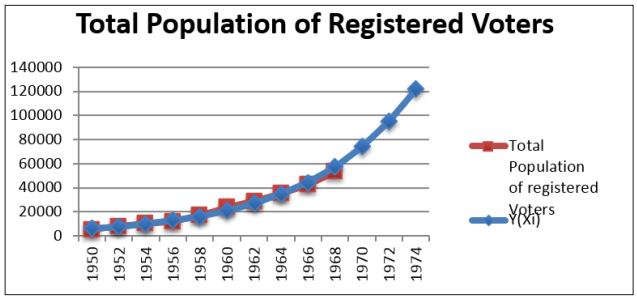

Figure 4: Comparison of Original Population of Registered Voters Data (y_i) and Theoretically Predictive Data, Y(X_i)

Equation (6) is the formula we obtained from our data analysis that can best describe the real-world Guam Registered Voters. We can now use Equation (6) to calculate the past and future Registered Voters. The predictive numeric values are listed in the following Table 3. In the next Step, we compute the error analysis

Table 3: Comparison of predicted and actual annul number of the Registered Voters in Guam

Year	1970	1972	1974
x_i	11	12	13
Predicted, $Y(x_i)$	73880.19	94987.45	122125
Actually achieved, y_i	to be verified	to be verified	to be verified

Step 10

Error analysis of the Least Square numeric algorithm.

Table 4 in the following show our statistical results and the errors between original given data and the Least Square Method theoretically predicted data [5]. We can see from the Figure 5 that direct error between original data and predicted data is very small and almost equally distributed at each data point. However the total Least Squared error is huge: 2563914.8 but error distributed along each data point is huge different. In order to make the minimum Least Squared total error become smaller, we can use a weighted factor to adjust the Least Squared error value at each data point [9].

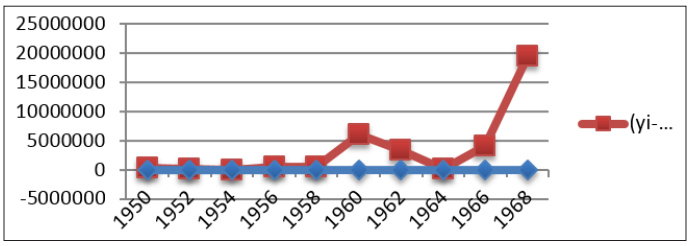

Figure 5: Comparison of Difference and Least Square Errors between Original Data and Predicted Data at Each Data Point

Table 4: Statistical Analysis of Errors between Industrial Data y_i and theoretically predicted Data $Y(x_i)$

Error Type	Error Formula	Total Error Value
Total Data Difference	$\sum_1^{10}(y_i - Y(x_i))$	898.3
Total Least Square	$Q(X)=\sum_1^{10}(y_i - Y(x_i))^2$	452343.65
Mean Value of Original Data	$\bar{y} = \frac{1}{n}\sum_1^n y_i$	2373.2
Total Difference of data and mean value	$\sum_1^{10}(y_i - \bar{y})$	0.0
Variance	$\delta_n^2 = \frac{1}{n}\sum_1^n(y_i - \bar{y})^2$	2563914.8
Standard Deviation	δ_n	1601.2
Correlational Coefficient	To be completed	To be done

Conclusion

In this paper, we used the total amount of registered voters on Guam as a case study, showed step by step the procedures on how to bind visualization and numeric methods together to analyze a massive amount of data in Guam. We used a mathematical transformation that turned the nonlinear data pattern problem into a linear data analysis problem by both using the visualization tool and numeric tool based on graphics showing the data properties. Then finally, we used the Least Square numeric methods to find a theoretical formula to describe and predict the future performance of the selected data. Readers will be able to understand our methods through the case study in the paper and use these effective methods to analyze real-world data with self-made computer programs and data analysis software.

References

1. Yadav C, Wang S, Kumar M (2013) "Algorithm and approaches to handle large data – a survey". International Journal of Computer Science and Network 2: 2277-5420.
2. Reichman OJ, Jones MB, Schildhaver MP (2011) Challenges and opportunities of open data in ecology. Science 331: 703-705.
3. 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.
4. 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.
5. 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.
6. 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.
7. Lichten W (1988) Data and error analysis. Allyn and Bacon, Inc 171.
8. Petras I, D Bednarova D (2010) Total Least Square approach to modeling: a Mat Lab toolbox. Acta Montanistica Slovaca 15: 158-170.
9. Bjorck A (1996) Numerical methods for Least Square problems. SIAM https://epubs.siam.org/doi/book/10.1137/1.9781611971484.
10. Wolberg J (2005) Data Analysis using the Method of Least Squares; extracting the most information from experiments. Springer https://link.springer.com/book/10.1007/3-540-31720-1.
11. Bureau of Statistics and Plans (2013) Guam Statistical Guam Statistical Yearbook https://bsp.guam.gov/wp-bsp-content/uploads/govarchieve/G11-30.101%202013_Guam%20Statistical%20Yearbook.pdf.