Research on the Development of Wordle Based on Multi-Model Analysis

Ziyue Zhang *

College of Science, Liaoning Technical University, Fuxin, Liaoning, 123000

* Corresponding Author Email: yuziyue777@126.com

Abstract. Wordle is becoming more and more popular around the world, so it is of great significance to study its healthy development. This paper predicts the number of Wordle 's report results and predicts the percentage of attempts for different words. Firstly, this paper establishes a prediction model of the number of ARIMA report results, and analyzes the relationship between word attributes and the percentage of players in difficult mode. This article collects data on the number of Wordle 's daily reporting results from January 7 to December 31, 2022, and predicts that the number of reporting results in 2023 will be: 10220-10637. This paper constructs three indicators to measure word attributes: the number of vowel letters, the number of affixes and the number of repeated letters. Using Pearson correlation coefficient method and AIC information criterion, according to the correlation coefficient of the three indicators, it is analyzed that there is no relationship between word attributes and the percentage of players in difficult mode. Then, this paper establishes a prediction model of the distribution of the number of attempts, and accurately predicts the percentage of attempts of EERTE words. This paper constructs the data into a lexicon and quantifies the letters. Multiple linear regression equation and multiple nonlinear regression equation were established by using the number of vowel letters, the number of affixes and the number of repeated letters corresponding to each letter in the word. The average number of guessing words and variance were fitted. It was found that the fitting effect of multiple nonlinear regression was better, R^2 was 0.805 and 0.821. Finally, the related attributes of “EERIE” were counted, and its distribution percentage was obtained. The results were 0, 1 %, 16 %, 49 %, 29 % and 3 %. The model constructed in this paper can provide some theoretical support for the good development of Wordle.

Keywords: Wordle; Word attributes; ARIMA; Pearson correlation coefficient; Regression analysis.

1. Introduction

Wordle is a popular software that provides a real five-letter English word as a puzzle every day. After selecting the regular mode or the difficult mode, the player will try to solve the problem six times or less, depending on the feedback they receive each time. Finally, players can share their completion time and guessing words on the social platform. Wordle [1] tops the list of the world 's most searched lists in 2022, and its popularity continues to grow.

In 1970, Box, G. E. P. and Jenkins, G. M. demonstrated in “Time Series Analysis: Forecasting and Control” how to select the parameters of the ARIMA model by observing the autocorrelation and partial autocorrelation of the data and use the model to predict future economic indicators. Chatfield, C. used the ARIMA model in 1975 to predict seasonal sales data, such as holiday shopping seasons or seasonal sales fluctuations in specific industries. By identifying and modeling seasonal components in time series, ARIMA models can be used to predict future sales trends. In 1991, Brockwell, P. and Davis, R. introduced how to use the ARIMA model to establish a time series model of infectious disease outbreaks in the book “Time Series: Theory and Methods”. By analyzing and modeling infectious disease data, we can predict the trend of disease transmission, grasp the dynamics of the epidemic, and formulate corresponding prevention and control strategies.

Based on the results of the statistical daily report from January 7 to December 31, 2022, this paper analyzes the documents and solves the following problems:

Develop a model that explains why the number of reported results varies from day to day and predicts the number of reported results on March 1, 2023.Discuss whether any of the attributes of the word affect the percentage of the player's score in hard mode, and why.

Build a prediction model and, for a given date of a puzzle, predict the percentage of attempts that will be one, two, three, four, five, six or fail to solve the puzzle, and list the uncertainties associated with the prediction model. Using the word EERIE on March 1, 2023, as an example, the results distribution reported on that day is predicted and the reliability of the prediction model in this paper is illustrated.

2. Materials and Methods

2.1. Data Acquisition and Preprocessing

The data used in this paper are derived from the C-question of the 2023 American Undergraduate Mathematical Contest in Modeling.(http://www.comap.com/undergraduate/contests/)In the data given in the original title, this paper finds that the spelling of the 314th word " tash " is wrong and changed to " trash " ; 525th word " clen ", misspelling changed to " clean " ; the number of reports on the day of the 529th word " study " is quite different from the number of reports on other dates, which is unreasonable, so this article deletes the data when calculating the problem ; the 545th word " rprobe " was incorrectly written and changed to " probe ". The data is normalized to obtain the percentage of difficult mode players ' scores.

2.2. Methods to Introduce

2.2.1 Arima Prediction Model

ARIMA model [2-4] is a time series analysis model proposed by American scholars Box and Jenkins in the 1920 s. It is one of the commonly used prediction methods. Among them, ARIMA (p, d, q) is a differential autoregressive moving average model, AR is an autoregressive, MA is a moving average, p, d, q is the number of autoregressive items, the number of moving average items and the number of differences, respectively. The basic idea of ARIMA model is to establish a model that can describe the characteristics of data through the transformation of autoregressive, moving average and difference of time series data, and use this model to predict future data changes. ARIMA model can better deal with the characteristics of many time series data, such as seasonality, trend, periodicity, etc., and can use fewer parameters to fit the data. The advantage of the ARIMA model is that the model is simple, and only endogenous variables are needed without the help of other exogenous variables. The ARIMA prediction flow chart is shown in Figure 1:

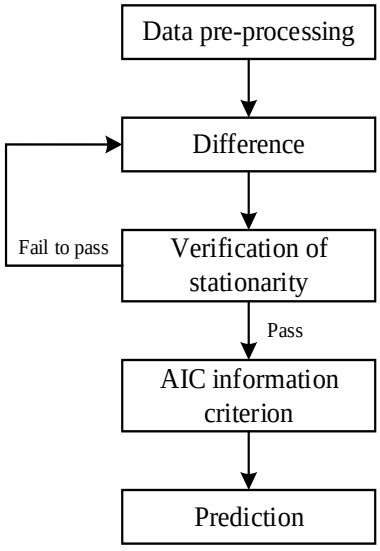

Figure 1: ARIMA forecasting model.

2.2.2 AIC information criterion

The AIC information criterion [5] is a standard to measure the goodness of fit of statistical models, also known as Akaike information criterion. It was founded and developed by Japanese statistician Akaike Hiroji. The AIC information criterion is based on the concept of entropy, which can weigh the complexity of the estimated model and the goodness of the model fitting data. The method of the Akaike information criterion is to find a model that can best explain the data but contains the least free parameters.

In general, AIC can be expressed as:

$$AIC = 2k - 2\ln(l) \tag{1}$$

Where k is the number of parameters and L is the likelihood function.

2.3. Model-Evaluation Index

In this paper, the complex determination coefficient R^2 [6] is used as the evaluation index of the model to measure the goodness of fit of the model. The formula is as follows:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum (Y_i - \hat{Y})^2}{\sum (Y_i - \bar{Y})^2} \tag{2}$$

3. Model Establishment and Solution

3.1. ARIMA prediction report result quantity model

3.1.1 Establish ARIMA prediction report result quantity model

Step1. The stationarity test of the original sequence is carried out. ARIMA model requires the sequence to meet the stability, so this paper obtains the ADF test results, including variables, difference order, T test results, AIC value, etc., to test whether the time series is stable**.** Table 1 is ADF test table.

Table 1: ADF test table

ADF test table				
variable	difference order	t	P	AIC
Number of reported results	0	-3.91	0.002***	7171.999
	1	-4.23	0.001***	7164.342
	2	-10.761	0.000***	7144.259

By analyzing the results of the ADF test [7], we get when the difference is 0, the significant P value is 0.002***, which is significant at the level, rejects the null hypothesis, and the sequence is a stable time series. When the difference is divided into 1 order, the significance P value is 0.001***, showing significance at the level, rejecting the null hypothesis, and the sequence is a stationary time series. When the difference is divided into 2 order and the significance P value is 0.000***, the level is significant, reject the null hypothesis, the sequence is a stable time series.

Step2. Determine the p and q values. Firstly, we looked at the data comparison chart before and after the difference, and found that it showed a stable trend. Then, we carried out autocorrelation analysis and partial autocorrelation analysis on the time series, and obtained ACF diagram and PACF diagram. We found that the autocorrelation diagram and partial autocorrelation diagram of the stationary sequence had tails, so the ARIMA model was established. Combined with the most significant order in the PACF and ACF plots, we estimate the p and q values as : 1 and 0. ACF graph and PACF graph are shown in Figure 2.

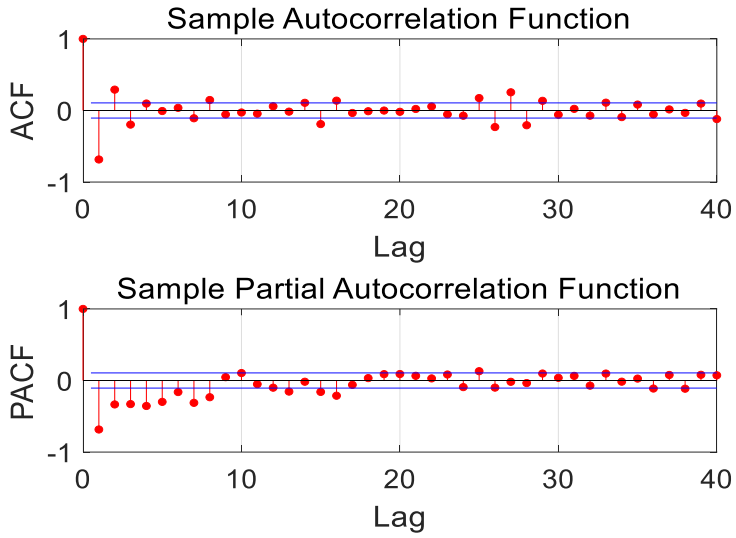

Figure 2: ACF graph and PACF graph

Step3. The test of model residual as white noise. ARIMA model requires the model to have pure randomness, that is, the model residual is white noise. Therefore, we look at the model test table and test the white noise of the model according to the P value of the Q statistic .Table 2 is Model check table.

Table 2: Model check table

Model Test Sheet			
	coefficient	standard deviation	t
constant	-168.768	465.243	-0.363
Number of reported results	-0.362	0.049	-7.322

According to the above table, it can be obtained from the analysis of Q statistics that Q6 does not show significance at the level, and the assumption that the residual of the model is a white noise sequence cannot be rejected. At the same time, the goodness of fit R^2 of the model is 0.982. The model performs well and the model basically meets the requirements.

Step4. Forecast according to the model formula and time series diagram. In order to analyze the model formula, we get the coefficient, standard deviation and T test results of the model. Based on the variable Number of reported results, we find the optimal parameters according to the AIC information criterion, and the model formula is shown in Figure 3.

$$y(t) = -168.768 - 0.3662 * y(t-1) \tag{3}$$

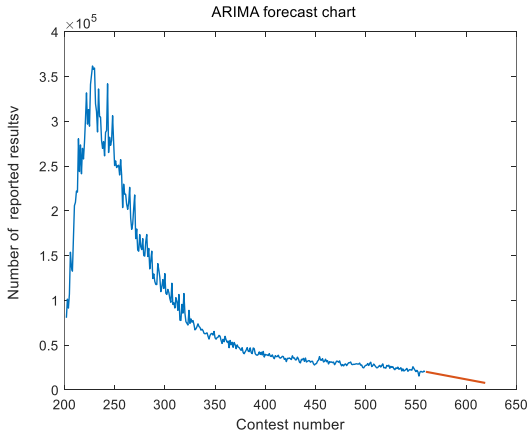

Figure 3: Time series prediction graph

Comprehensive analysis combined with time series analysis chart to obtain backward prediction order results.

The interval of the number of players is constructed with the upper and lower limits of 2 % error. Finally, from the model formula and the time series diagram, we predict that the number of reported results on March 1,2023 is : 10220-10637.

3.1.2 Pearson correlation coefficient analysis

First of all, the number of cause letters and affixes are important components of English words. The mastery of the number of cause letters and affixes is more conducive to our memorizing words, and it is easier to guess; secondly, some words are composed of multiple repeated letters. If it is found that the word is composed of multiple repeated letters in the word guessing game, it will be easier to guess the puzzle if the range of guessing words is reduced. Therefore, we choose the number of cause letters, the combination of affixes and the number of identical letters in an English word as indicators to measure word attributes. Then we processed the data in the attachment, noting the number of cause letters in each word, the number of prefixes and suffixes, and the number of identical letters in an English word. In order to find out the relationship between the word attributes and the percentage of players ' scores in difficult mode, we choose Pearson correlation coefficient method for analysis.

Pearson correlation coefficient [8] is a linear correlation coefficient, which reflects the linear correlation degree of the two variables. The value is between-1 and 1, and the closer the absolute value is to 1, the stronger the linear correlation intensity of the two variables is.

This topic calculates the correlation coefficient between the word attribute and the percentage of players ' scores in difficult mode. First of all, let X_i, Y two index variables, X_i are the word attribute, X_1 is the number of letters, X_2 is the combination of prefix and suffix, X_3 is the number of the same letters in an English word, Y is the percentage of the player 's score in the difficult mode, $\rho_{X_i,Y}$ represents the Pearson correlation coefficient between the two indexes, then the Pearson correlation coefficient between the two index variables is defined as:

$$\rho_{X_i,Y} = \frac{\operatorname{cov}(X_i,Y)}{\delta_{X_i}\delta_Y} = \frac{E\left[\left(X_i-\mu_{X_i}\right)\left(Y-\mu_Y\right)\right]}{\delta_{X_i}\delta_Y} \tag{4}$$

Among them, $\operatorname{cov}(X_i,Y)$ is the covariance between the index i and Y, δ_{X_i} is the overall standard deviation of the index i variable, μ_{X_i} is the overall mean of the index i variable.

According to our statistical data and the percentage of players scoring in difficult mode, we use MATLAB to get the correlation coefficient table, and we draw the table into the correlation coefficient heat map, Pearson correlation coefficient heat map is shown in Figure 4.

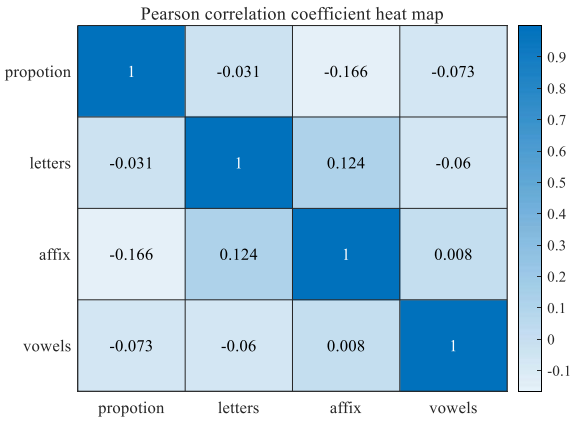

Figure 4: Pearson correlation coefficient heat map

From the graph, we can see that the absolute value of the correlation coefficient between the word attribute and the percentage of players ' scores in difficult mode is 0-0.2, so we determine that the word attribute is not related to the percentage of players ' scores in difficult mode. Moreover, Wordle 's game rule is to select the mode first and then solve the puzzle. Therefore, we believe that word attributes do not affect the percentage of player scores in the difficult mode.

3.2. Prediction model of attempt times distribution

3.2.1 Establish a distribution prediction model for the number of attempts

Step1. Using the given data, the proportional distribution maps of the scores of the players after trying once to trying X times for each word are drawn respectively. It is found that the scores are approximately in accordance with the normal distribution. The proportional distribution maps of the four words are shown in Figure 5.

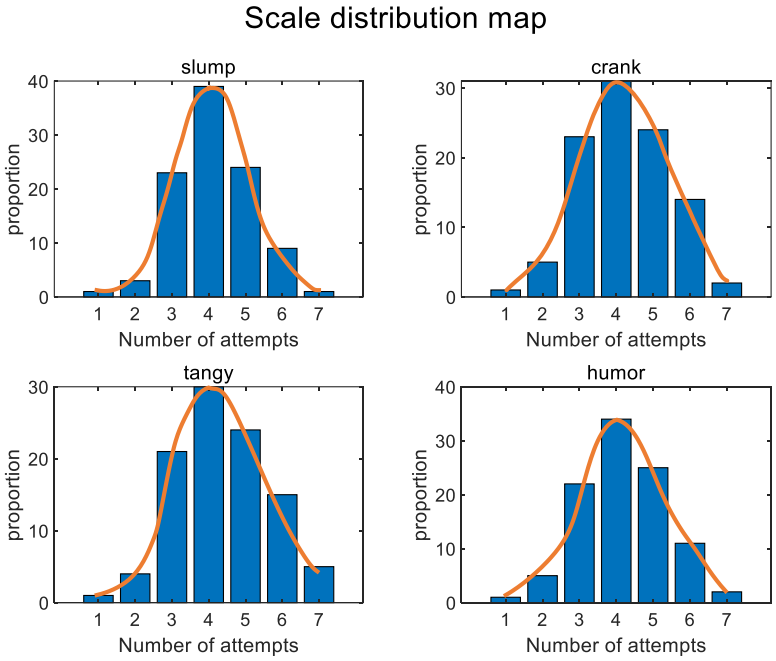

Figure 5: Scale distribution map

Step2. The mean and variance of the number of attempts by players for each word from January 7, 2022 to December 31, 2022 are calculated. Therefore, the mean value can reflect the difficulty of words to a certain extent. The greater the mean value, the higher the difficulty of words, and the more difficult it is to guess.

Step3. The words after data processing are split, the five letters are split, and the number of occurrences of each letter in each word in this thesaurus is counted respectively, and the order is sorted from high to low.

tep4. Encode from high to low according to the frequency of each letter. The letter e with the highest frequency is coded as 26, and the letter j with the lowest frequency is coded as 1, and so on, and the matrix is constructed. Therefore, coding can reflect the common degree of letters. The larger the coding is, the more common it is, and the easier it is to guess.

Step5. According to the letter coding in the previous step, combining the three indicators used to measure the word attributes in the first question: the number of cause letters, the combination of affixes and the number of the same letters in an English word, we decided to use the multiple linear regression model to calculate the relationship between the two variables.

Multivariate regression analysis theory is an important theory in statistics. Based on the correlation theory, it is a mathematical statistics method used to deal with the correlation between variables. It is generally used to study the internal relationship between a random variable Z and multiple variables

$X_1 \sim X_i$, and to analyze its internal changes. The key of multiple regression analysis is to design regression model.

According to the theory of multiple regression, we first use multiple linear regression model, the expression is as follows:

$$Z = \beta_0 + \sum_{i=1}^{8} \beta_i X_i \tag{5}$$

Step5. According to the letter coding in the previous step, combining the three indicators used to measure word attributes in the first question: the number of cause letters, the combination of affixes, and the number of letters appearing in an English word, we decided to use the multiple linear regression model [9]to calculate the relationship between the two variables.

Multivariate regression analysis theory is an important theory in statistics. Based on the correlation theory, it is a mathematical statistics method used to deal with the correlation between variables. It is generally used to study the internal relationship between a random variable Z and multiple variables $X_1 \sim X_i$, and to analyze its internal changes. The key of multiple regression analysis is to design regression model.

According to multiple regression theory, we first use multiple linear regression model, the expression is as follows:

$$Z = \beta_0 + \sum_{i=1}^{8} \beta_i X_i \tag{6}$$

Where Z is the mean and variance, X_1 is the number of vowel letters, X_2 is the affix, X_3 is the number of repeated letters in a word, $X_4 \sim X_8$ represents the encoding of five letters in a word, $\beta_0 \sim \beta_8$ is the regression coefficient.

The data is imported into Matlab for statistical calculation. The results show that: $R^2 = 0.184$, indicating that the model equation can only explain 18.4 % of the data regression law. Therefore, the multiple linear regression model cannot fit the relationship between the eight indicator variables and the mean and variance. Therefore, according to the theory of multiple regression, we use the multiple nonlinear regression model [10] to calculate again, and the expression is as follows:

$$Z = \sum_{i=1}^{8} a_i X_i + \sum_{i=1}^{8} b_i X_i^{\,2} + \sum_{j=i+1}^{8} \sum_{i=1}^{8} c_{ij} X_i X_j + d \tag{7}$$

where a_i is the square coefficient; b_i is the interaction term coefficient ; c_{ij} is the coefficient of linear term ; d is a constant term

The values of the independent variable and the dependent variable are brought into the formula (4), and the coefficient of the equation is solved by MATLAB.

The following are the function coefficients fitted by the variance:

The square coefficient A is:

$$A = (a_1 a_2 \cdots a_8) = (-0.023 \quad -0.085 \quad -0.016 \quad -0.005 \quad -0.031 \quad -0.056 \quad 0.344 \quad 0.072) \tag{8}$$

The interaction term coefficient B is:

$$B = (b_1 b_2 \cdots b_8) = (0 \quad 0.001 \quad 0.001 \quad 0 \quad 0.001 \quad -0.034 \quad -0.011 \quad 0.025) \tag{9}$$

The monomial coefficient C is:

$$C=\begin{pmatrix} c_{12} & c_{13} & c_{14} & c_{15} & c_{16} & c_{17} & c_{18} \\ & c_{23} & c_{24} & c_{25} & c_{26} & c_{27} & c_{28} \\ & & c_{34} & c_{35} & c_{36} & c_{37} & c_{38} \\ & & & c_{45} & c_{46} & c_{47} & c_{48} \\ & & & & c_{56} & c_{57} & c_{58} \\ & & & & & c_{67} & c_{68} \\ & & & & & & c_{78} \end{pmatrix}=\begin{pmatrix} 0.001 & 0 & 0 & 0 & 0.003 & 0.002 & 0.005 \\ & 0.001 & 0 & 0 & 0.002 & -0.003 & 0.006 \\ & & -0.001 & 0 & 0.004 & -0.001 & -0.006 \\ & & & 0 & 0.002 & -0.003 & -0.001 \\ & & & & 0.005 & -0.006 & -0.003 \\ & & & & & -0.037 & -0.079 \\ & & & & & & 0.035 \end{pmatrix} \quad (10)$$

The constant term coefficient D is:

$$D=2.399 \quad (11)$$

Similarly, the function coefficient fitted by the mean value can be obtained.

The values of independent variables and dependent variables are imported into Matlab again for statistics. The calculation results show that $R^2=0.805$, indicating that the fitting degree of the model equation is high, which can explain 80.5 % of the data regression rate, and 19.5 % of the factors are uncontrollable factors. The conclusion is that the mean value is 1.452494094 and the variance is 4.179708822.

Step6. According to the code corresponding to each letter of EERIE and three attribute indexes, the relationship between the mean, variance and eight attributes in the previous step is used to predict the mean and variance, and the distribution function is obtained. The proportion distribution map of the score after the player tries once to try X times for the word EERIE is drawn. Through its distribution map, the report results of EERIE on March 1,2023 are obtained. The distribution of the scores of the players who try 1 time, 2 times, 3 times, 4 times, 5 times, 6 times and more than 6 times is 0,1 %, 16 %, 49 %, 29 %, 3 % and 0, respectively, as shown in the following figure 6.

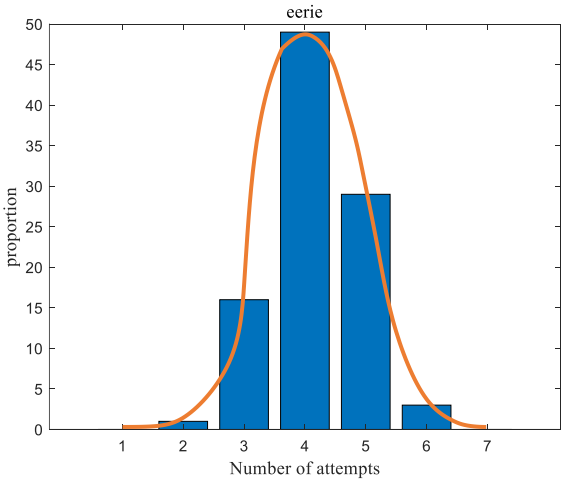

Figure 6: EERIE's map of reporting results

3.2.2 Uncertainty and model feasibility

The uncertainty of this topic includes: because this article is based on the data given by the topic, the vocabulary is limited, and the capacity of the vocabulary will affect the encoding of the letters, which in turn affects the calculation of the distribution function and affects the final result. The influence of factors such as part of speech and meaning on the subjective degree cannot be reflected in the established model. That is, the prediction effect will be biased. This article believes that the maximum number of guesses is 7 times, and the actual situation will differ.

The multivariate nonlinear regression model established in this paper R^2 is 0.805 and 0.821 respectively when fitting the average guessing times and variance. The degree of fitting is good, and the prediction reliability is high.

4. Conclusion

This paper first establishes the ARIMA prediction report result quantity model, and predicts that the number interval of the report on March 1,2023 is 10220-10637. Pearson correlation coefficient analysis was performed on the percentage of players scoring in the word attribute and the difficult mode, and the absolute value of the correlation coefficient was in the range of 0-0.2, indicating that the word attribute was not related to the percentage of players scoring in the difficult mode. Therefore, we believe that word attributes do not affect the percentage of player scores in difficult mode. Secondly, this paper establishes a prediction model for establishing the distribution of the number of attempts, and uses multiple linear regression equations and multiple nonlinear regression equations to fit the distribution function. It is found that the fitting effect of multiple nonlinear regression is better, which R^2 is 0.085 and 0.821. The distribution of the reported results of "EERIE" was predicted to be 0, 1 %, 16 %, 49 %, 29 % and 3 % by this model. In the future, this article hopes to quote more data to optimize the model, which is more conducive to the good development of Wordle.

References

[1] Match S. The New York Times buys Wordle [J]. The New York Times, 2022.

[2] Umemura S, Arima H, Arima S, et al. The Japanese Society of Hypertension guidelines for the management of hypertension (JSH 2019) [J]. Hypertension Research, 2019, 42(9): 1235-1481.

[3] Abonazel M R, Abd-Elftah A I. Forecasting Egyptian GDP using ARIMA models [J]. Reports on Economics and Finance, 2019, 5(1): 35-47.

[4] Katoch R, Sidhu A. An application of ARIMA model to forecast the dynamics of COVID-19 epidemic in India [J]. Global Business Review, 2021: 0972150920988653.

[5] Cavanaugh J E, Neath A A. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements [J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2019, 11(3): e1460.

[6] Wang H, Gao X, Lv Z, et al. Recent advances in oxidative R1-H/R2-H cross-coupling with hydrogen evolution via photo-/electrochemistry: focus review [J]. Chemical reviews, 2019, 119(12): 6769-6787.

[7] Yaya O O S, Ogbonna A E, Mudida R. Hysteresis of unemployment rates in Africa: new findings from Fourier ADF test [J]. Quality & Quantity, 2019, 53(6): 2781-2795.

[8] Armstrong R A. Should Pearson's correlation coefficient be avoided [J]. Ophthalmic and Physiological Optics, 2019, 39(5): 316-327.

[9] Etemadi S, Khashei M. Etemadi multiple linear regression [J]. Measurement, 2021, 186: 110080.

[10] Fan C, Ding Y. Cooling load prediction and optimal operation of HVAC systems using a multiple nonlinear regression model [J]. Energy and Buildings, 2019, 197: 7-17.