ISSN 2616-5775 Vol. 6, Issue 10: 100-105, DOI: 10.25236/AJCIS.2023.061015

Analysis of Wordle's Data Based on a Stepwise Regression Iterative Prediction Model

Tong Shi*, Yuxuan Zhao

Department of Applied Statistics, Anhui University, Anhui, Hefei, China
**Corresponding author: 1468662015@qq.com*

Abstract: *Wordle is currently a popular puzzle game featured daily in the New York Times. Players are required to guess a five-letter word in up to six attempts to solve the puzzle. This paper considers 30 word attributes that affect the percentage. It assigns values to the attributes by means of dummy variables and other methods in order to study the percentage of the number of players who succeed in solving the puzzle at different number of attempts. A stepwise regression model is established to determine the equation of the attributes affecting each percentage. It is found that the number of repeated letters in a word has the greatest impact on the difficulty of guessing the word. Finally, the word EERIE is used as an example for prediction analysis, which is predicted as a difficult puzzle.*

Keywords: *Stepwise Regression, Regression Equation, F-test, Dummy Variable*

1. Introduction

Wordle [1] is a puzzle game with only six tries per day. Due to its special game mechanics, it attracts many players to try it. Many users report on Twitter the number of tries when they succeed in guessing correctly. The results reported by the statistics include the words guessed and the percentage of those guessed on the day in between one and six tries or inability to solve the puzzles, respectively. Further research based on the reported results will be able to rationally explain the variation in the reported results and find the patterns that exist behind the fun game. At the same time, predictions are made about the results of future puzzles given to avoid the game being too difficult or too easy.

In the past years, fewer scholars have studied Wordle as a game. K Koh et al. (2010) introduced ManiWordle, a Wordle-based visualization tool that revamps interactions with the layout by supporting custom manipulations [2]. M Bonthron (2022) proposed combining a rank one approximation and latent semantic indexing to a matrix representing the list of all possible solutions [3]. Scholars at home and abroad have applied stepwise regression models in different fields. Chen et al. (2001) used the stepwise multiple regression statistical method calculate the relation between algal chlorophyll a (Chla), total algal biomass (TB), Microcystis biomass (MB) and these environmental factors [4]. Zhang et al. (2007) multi-objective optimization of automobile safety based on stepwise regression model [5]. Scholars at home and abroad have applied stepwise regression models in different fields. KVD Borght et al.(2011) applied a method based on stepwise regression to lower the complexity of these phenotype prediction models using a 3-fold cross-validated selection of mutations[6].Breaux and Harold J. (1968) modified Efroymson's technique for stepwise regression analysis[7].A Elzamly and B Hussin (2014) mitigated software maintenance project risks with stepwise regression analysis techniques [8].

In previous applications of stepwise regression models, there has been no application to the field of gaming and to better understand the current popular fun game. This paper analyses the current emergence of the game and makes predictions for the future based on the stepwise regression model. Firstly, we consider the factors affecting the proportion of tries when playing Wordle, quantify the degree of influence of the factors, and then establish a stepwise regression model to find the regression equations for the proportion of people who succeeded in each of the tries, and finally predict the percentage of tries using the word "EERIE" as an example.

ISSN 2616-5775 Vol. 6, Issue 10: 100-105, DOI: 10.25236/AJCIS.2023.061015

2. Preliminary

2.1 Variables representing the property

Dummy variables of whether the letters a/b/c... y/z appear in the words were introduced respectively, for a total of 26 new variables. The effects of the four factors of letter (values defined by the frequency of occurrence of commonly used letters), repeat (number of repeated letters), combination (values defined by the frequency of occurrence of commonly used two-letter combination in the data set) and daily (levels of frequency of daily use of a word) on the regression equation were also considered.

2.2 Data sources and processing

Data set from Wordle Game public data from 7 January 2022 to 31 December 2022 for daily results including date, match number, word of the day, and percentage of guesses in 1 try, 2 tries, 3 tries,4 tries, 5 tries,6 tries, or unable to solve the puzzle (X tries).

Variable called *repeat* represented that the same letter appears more than once between some words. This situation may make it more difficult for players to guess the words. Therefore, the words were quantified according to the number of repeated letters, and words without repeated letters were recorded as 0, words with one repeated letter were recorded as 1, and words with two repeated letters were recorded as 2.

There are many words with similar spelling available for guessing which contains the same double letter combination. By counting all the words given in the data set, we can get some two-letter combination with a high number of occurrences, and the number of occurrences of these two-letter combination will be counted and then quantified based on their next. The two-letter combination with less than 10 occurrences are recorded as 0, the two-letter combination with frequencies of 10 to 14 occurrences are recorded as 1, the two-letter combination with frequencies of 15 to 19 occurrences are recorded as 2, and the two-letter combination with frequencies of 20 or more occurrences are recorded as 3. Then, according to the spelling of the words, variable called combination represented by their recorded numbers of two-letter combination are added up to get the final quantified results. Table 1 illustrates the record of common two-letter combination in words.

Table 1: Record of common two-letter combination in words

common two-letter combination in words	record
an,ap,at,ca,ea,el,ha,ho,la,ll,me,mo,ne,ou,ra,re,ri,ro,se,te,to,tr,un,ve	1
ar,ch,or,th	2
al,er,in,st	3

Variable called *daily* represented levels of frequency of daily use of a word. The Collins dictionary divides vocabulary frequency into five dimensions, between most and least used. We define 1 as most frequently used, 5 as rarely used, and 2,3,4 as somewhere in between.

According to the dictionary summary, each letter appears differently in words. Some letters appear very frequently while others appear very infrequently. Therefore, letters can be assigned weights based on their frequency of occurrence, which represented the variable called *letter*. For example, there are nine letters that occur more than 5% of the time, and they are assigned weights according to their frequency. The letter e is used most frequently, so it is assigned a weight of 9, and the letter t is used second most frequently, so it is assigned a weight of 8... The rest of the letters that occur less than 5% of the time are assigned a weight of 0. Then the weights of the different letters in the word are added together to get the quantified value of the word. It is shown in Table 2.

Table 2: Weight of letters

letter	weight
e	9
t	8
a	7
o	6
n	4
r	1
…	…

ISSN 2616-5775 Vol. 6, Issue 10: 100-105, DOI: 10.25236/AJCIS.2023.061015

2.3 Stepwise regression

Step1: For the introduced n attribute independent variables $A_1, A_2 \ldots A_k (k = 1,2, \ldots, n)$, establish the multivariate linear equation.

$$y_X = \beta_0 + \beta_1 A_1 + \beta_2 A_2 + \beta_3 A_3 + \cdots + \beta_k A_k \quad (1)$$

y_X is percent in X try/tries($X = 1,2,3,\ldots 6,7$ or more). Calculate the values of the F-test statistics of the corresponding regression coefficients of the attribute variables $A_1, A_2 \ldots A_k$respectively, denoted as $F_1^{(1)}, F_2^{(1)} \ldots F_k^{(1)}$, and take the maximum of them as $F_{i1}^{(1)}$, i.e. $F_{i1}^{(1)} = max\{F_1^{(1)}, F_2^{(1)} \ldots F_k^{(1)}\}$, For a given significance level, Denote its corresponding critical value as $F^{(1)}, F_{i1}^{(1)} \geq F^{(1)}$, Introduce A_{i1} into the regression model.

Step2: A binary regression[9] model between the dependent variable per cent in X try/tries and the subset of attribute independent variables $\{A_{i1}, A_1\}, \{A_{i1}, A_2\}, \ldots, \{A_{i1}, A_n\}$ between a total of 29 binary regression models. Calculate the value of the statistic for the regression coefficient F test of the variable, denoted as $F_k^{(2)}$, The largest item is recorded as $F_{i2}^{(2)}$, i.e. $F_{i2}^{(2)} = max\{F_1^{(2)}, F_2^{(2)} \ldots F_k^{(2)}\}$。 For a given significance level α,its corresponding critical value is noted as $F^{(2)}$ If $F_{i2}^{(1)} \geq F^{(2)}$, A_{i2} is introduced into the regression model, otherwise the variable is not introduced.

Step3: the regression of the dependent variable on a subset of the variables is considered and step2 is repeated. By repeating this process, one of the independent variables not introduced in the regression model is selected each time until no variables are introduced by the test.

3. Experiments

3.1 Analysis of experimental results

A stepwise regression of each dependent variable per cent in X try/tries was performed to obtain the set of their reserved variables separately. Table 3 shows the specific variables.

Table 3: Set of reserved variables

Percent in	D: Set of reserved variables
1 try	*letter, daily, repeat*
2 tries	*letter, repeat, daily, e, w, x, o, r, v, y*
3 tries	*repeat, t, w, daily, z, y, p, x, g, v, f, j, i, k, m, q, e, letter*
4 tries	*p, z, combination, n*
5 tries	*repeat, t, combination, z, w, daily, p, x, v, g, i, f, j, q, y, l, e, b, m, k*
6 tries	*repeat, t, z, y, p, w, daily, v, j, x, g*
7 or more tries (*X*)	*repeat, combination, y, c*

Table 4: Stepwise regression model results

Variable	Non-standardised coefficient		Standardised coefficient	P	VIF	R^2	Adjusted R^2	F
	B	Standard error	Beta					
constant	33.62	1.698	0	0.000***	-	0.843	0.818	F=52.147P=0.000***
repeat	-5.03	0.536	-0.354	0.000***	1.046			
t	1.099	0.822	0.075	0.182	2.287			
w	-5.971	1.067	-0.214	0.000***	1.079			
daily	-1.346	0.275	-0.186	0.000***	1.061			
. . .								
letter	0.187	0.086	0.159	0.030**	3.878			

We can observe that the variables *repeat, daily, letter, and combination* were more commonly retained during the seven stepwise regressions conducted, with the variable *repeat* being retained the most often and introduced with a higher significance, indicating that the variable *repeat* has a greater impact on the percentage of reported results. It is noteworthy that the variable *letter* was retained in all of the first three stepwise regressions, and the variable *combination* was retained to appear in the last

ISSN 2616-5775 Vol. 6, Issue 10: 100-105, DOI: 10.25236/AJCIS.2023.061015

four regressions. We can find that for the first few attempts, people prefer to choose words made up of common letters when useful information is lacking. For the last few attempts, when information about letters that would not appear in the correct word was obtained, people were more willing to form words and fill in words by choosing letters that might appear and their letter combinations. The table of stepwise regression model results are shown in Table 4.

It can be obtained that the significance *P-value* is 0.000***, and the level presents significance, rejecting the original hypothesis that the regression coefficient is 0. For the performance of whether there is co-linearity in the variables, VIF is all less than 10, so the model has no problem of multiple co-linearity and the model is well constructed. The real value and the predicted value in the graph of fitting effect change the same trend with small difference, meanwhile, the indicator R^2in the table is close to 0.85, and the model fit is good. The equation of the stepwise regression model on the dependent variable per cent in 3 tries is as follows:

$$y = 33.62 - 5.03 \times repeat + 1.099 \times t \ldots - 2.151 \times e + 0.187 \times letter \quad (2)$$

A stepwise regression method for the percentage of remaining tries obtained the following results, shown in Table 5.

Table 5: Regression equations of percent in X tries

Dependent Variable	Equation of the model
Percent in 1 try	$y_1 = 0.788 + 0.022 \times \text{letter} \ldots - 0.229 \times \text{repeat}$
Percent in 2 tries	$y_2 = 7.363 + 0.292 \times \text{letter} \ldots - 0.964 \times y$
Percent in 3 tries	$y_3 = 33.62 - 5.03 \times \text{repeat} \ldots + 0.187 \times \text{letter}$
Percent in 4 tries	$y_4 = 33.267 + 2.518 \times p \ldots + 1.358 \times n$
Percent in 5 tries	$y_5 = 13.784 + 3.551 \times \text{repeat} \ldots + 1.687 \times k$
Percent in 6 tries	$y_6 = 4.267 + 3.185 \times \text{repeat} \ldots + 1.467 \times g$
Percent in 7 tries or more	$y_7 = 0.08 + 1.215 * \text{repeat} \ldots + 1.044 \times c$

3.2 Test of stepwise regression

The F-tests were performed on the seven stepwise regression models and their goodness-of-fit to the observations were calculated as Table 6 shows.

Table 6: Regression model test results

Indicator		Percent in						
		1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (*X*)
F-test	F	22.591	44.529	52.147	48.914	51.758	47.66	25.837
	P	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
R^2		0.397	0.716	0.843	0.792	0.866	0.762	0.463
Adjusted R^2		0.389	0.699	0.818	0.782	0.84	0.741	0.452
Note: ***, **, * represent 1%, 5%, 10% level of significance respectively								

The table shows that all seven stepwise regression equations pass the test and are significant, indicating that the stepwise regression model is suitable for the prediction of this problem. When the dependent variable is percent in *X* tries (*X*=2,3,4,5,6), the$R^2$of the fit is above 0.7 and the fit is excellent. However, when the dependent variable is percent in 1 try, the fit is poor.

4. Analytical forecast of the example EERIE

The specific word EERIE is given in the question for us to make predictions. First, we need to obtain the non-zero indicators for each attribute of the word EERI. Table 7 shows the attributes of EERIE.

Table 7: Attributes of EERIE

Attributes of EERIE			
Attribute	Value	Attribute	Value
e	3	*r*	1
i	1	*repeat*	3
daily	4	*letter*	33
combination	4		

ISSN 2616-5775 Vol. 6, Issue 10: 100-105, DOI: 10.25236/AJCIS.2023.061015

Then, the corresponding values are brought into the equation of the stepwise regression model on the dependent variable per cent in 3 tries to obtain the predicted value $\hat{y} = 14.117$. Similarly, the equations of the regression models of the remaining dependent variables can be obtained separately and brought into the attribute data of EERIE for prediction, and the specific results are shown in Table 8.

Table 8: Prediction results

Dependent Variable	Equation of the model	Predicted results
Percent in 1 try	$y_1 = 0.788 + 0.022 \times letter \ldots - 0.229 \times repeat$	$\widehat{y_1} = 0.351$
Percent in 2 tries	$y_2 = 7.363 + 0.292 \times letter \ldots - 0.964 \times y$	$\widehat{y_2}$ =3.846
Percent in 3 tries	$y_3 = 33.62 - 5.03 \times repeat \ldots + 0.187 \times letter$	$\widehat{y_3}$ =14.117
Percent in 4 tries	$y_4 = 33.267 + 2.518 \times p \ldots + 1.358 \times n$	$\widehat{y_4}$ =31.147
Percent in 5 tries	$y_5 = 13.784 + 3.551 \times repeat \ldots + 1.687 \times k$	$\widehat{y_5}$ =29.154
Percent in 6 tries	$y_6 = 4.267 + 3.185 \times repeat \ldots + 1.467 \times g$	$\widehat{y_6}$ =30.085
Percent in 7 tries or more	$y_7 = 0.08 + 1.215 * repeat \ldots + 1.044 \times c$	$\widehat{y_7}$ =5.265

Finally, the predicted results are normalized and the data are rounded to obtain the predicted percentages of (1, 2, 3, 4, 5, 6, *X*) for a future date is (0, 3, 12, 26, 27, 27, 5).

5. Conclusions

The number of repeated letters between words, the frequency of common affixes in words, the frequency of common letters in words, and the degree of common use of words, which affect the percentage of report results, are set as variables, and then introduce whether the letters *a/b* appear in the words. A total of 30 variables of dummy variables of *y/z* are introduced into the model. An F test is performed each time a variable is introduced, and a T test is performed on each of them. In the seven regression, we can observe that the variables *repeat, daily, letter, and combination* were more commonly retained during the seven stepwise regressions conducted. So, the equation of the stepwise regression model on the dependent variable per cent in 3 tries is $y = 33.62 - 5.03 \times repeat + 1.099 \times t \ldots - 2.151 \times e + 0.187 \times letter$.

And F-test shows that all seven stepwise regression equations pass the test and are significant, indicating that the stepwise regression model is suitable for the prediction of this problem. When the dependent variable is percent in *X* tries (*X*=2,3,4,5,6), the R^2 of the fit is above 0.7 and the fit is excellent. However, when the dependent variable is percent in 1 try, the fit is poor.

For the specific word EERIE, the corresponding values are brought into the equation of the stepwise regression model on the dependent variable per cent in 3 tries to obtain the predicted value $\hat{y} = 14.117$.

Finally, the predicted results are normalized and the data are rounded to obtain the predicted percentages of (1, 2, 3, 4, 5, 6, *X*) for a future date is (0, 3, 12, 26, 27, 27, 5).

References

[1] Anderson B J, Meyer J G. Finding the optimal human strategy for Wordle using maximum correct letter probabilities and reinforcement learning [J]. 2022.

[2] Elzamly, Abdelrafe, and B. Hussin. Mitigating Software Maintenance Project Risks with Stepwise Regression Analysis Techniques. Journal of Modern Mathematics Frontier 3. 2(2014):34-44.

[3] Breaux, and J. Harold. A modification of Efroymson's technique for stepwise regression analysis. Communications of the Acm 11. 8(1968):556-558.

[4] Borght, Koen Van Der, et al. Cross-validated stepwise regression for identification of novel non-nucleoside reverse transcriptase inhibitor resistance associated mutations. BMC Bioinformatics, 12, 1 (2011-10-03) 12. 1(2011):386.

[5] Wan D, Wang Y, Gu N, et al. A novel approach to extreme rainfall prediction based on data mining[C]//International Conference on Computer Science & Network Technology.IEEE, 2012. DOI:10.1109/ICCSNT.2012.6526285.

[6] Chen, Y. W., B. Q. Qin, and X. Y. Gao. Prediction of Blue-green Algae Bloom Using Stepwise Multiple Regression between Algae & Related Environmental Factors in Meiliang Bay, Lake Taihu. Journal of Lakeence 13. 1(2001):63-71.

[7] Bonthron, Michael. Rank One Approximation as a Strategy for Wordle. arXiv e-prints (2022).

ISSN 2616-5775 Vol. 6, Issue 10: 100-105, DOI: 10.25236/AJCIS.2023.061015

[8] Koh, Kyle, et al. Mani Wordle: Providing Flexible Control over Wordle. IEEE Transactions on Visualization and Computer Graphics 16. 6(2010):1190-1197.
[9] Wu Y, Zhang Q, Hu Y, et al. Novel binary logistic regression model based on feature transformation of XGBoost for type 2 Diabetes Mellitus prediction in healthcare systems[J]. Future generations computer systems: FGCS, 2022:129.