Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 5, No. 1, 2023

Application and Research based on ANOVA and Logistic Regression Models

Guowei Li, Shanwei Yang, Sai Li, Jie Jian, Juan Li *

School of information Engineering, Wuhan Business University, Wuhan 430056, China
*** Corresponding author:** Juan Li (Email: looj@wbu.edu.cn)

Abstract: The Silk Road was a passage for cultural exchanges between China and the West in ancient times, of which glass was valuable material evidence of early trade exchanges, and China's early glass also led to different chemical compositions after absorbing some foreign technologies. For example, nowadays, most glass cultural relics are divided into two categories, and the identification, analysis and classification of them and similar problems are also the direction that needs to be studied. In view of such problems, we propose to solve such problems by studying one-way ANOVA and binary classification logistic regression model algorithm, in which one-way ANOVA is a significant test of the mean difference between two or more samples (i.e. factors), that is, it performs certain screening and testing on the relevant data of the required classification identification, and then completes the classification and identification of the results of the screening test by the 0-1 variable and the dependent variable of the two-way logistic regression model. The test of the result prediction by extracting the differential data obtained by the former, that is, its sensitivity analysis, can confirm the accuracy and effectiveness of the two model algorithms in this regard.

Keywords: 0ne-way ANOVA; Dichotomous Logistic Regression; Sensitivity Analysis.

1. Introduction

At present, with the archaeological research and discovery of cultural relics in China, a large number of archaeological research experts have joined in, and among them, the identification and historical cognition and analysis of a large number of cultural relics after their discovery is the focus and difficulty of the work afterwards, and in the process of identifying cultural relics and types and components, a large set of influencing factors and experimental controls and natural influences lead to hindrance also make the process more tedious.

One-factor ANOVA is a statistical test that compares the difference between the means of groups in a sample when only one independent variable or factor is considered. Based on this, Ni Feng makes it widely used in the homogeneity test of samples by explaining the principle and calculation steps of factor ANOVA and combining it with practical, shallow analysis of one-factor ANOVA in the homogeneity test of coal samples. Binary logistic regression model is also one of the most classical methods of machine learning. Based on this, Wu Jiao established a binary logistic regression model for studying the factors influencing the nutritional status of children aged 1 to 3 years old, which enables him to take appropriate intervention strategies in advance to prevent malnutrition in children and improve their nutritional status. Based on the extensive research and application of these models and algorithms in recent years, we intend to use these models and algorithms to assist in solving the problem of identifying glass artifacts, so that the difficulty of their identification can be further alleviated and helped.

2. Problem Description

The main raw material of glass is quartz sand, the main chemical composition of which is SiO2, and due to the high melting point of pure quartz sand, fluxes are added during refining to lower the melting temperature. The fluxes commonly used in ancient times were grass ash, natural alkali, saltpeter and lead ore, etc. Limestone was added as a stabilizer, which was converted to CaO after calcination. the main chemical composition of the fluxes added differed. For example, lead barium glass adds lead ore as a flux in the firing process, and its content of PbO and BaO is high, which is usually considered as our own invented glass species, and the glass of Chu culture is mainly lead barium glass. Potassium glass is made by firing substances with high potassium content, such as grass wood ash, as fluxes, and is mainly popular in Lingnan, China, and other regions in Southeast Asia and India. In this context, we intend to analyze the chemical composition of an existing group of glass artifacts of unknown category, identify the type to which they belong, and analyze the sensitivity of the classification results.

3. Modeling and Solving Problem 2

3.1. Screening Data by ANOVA

Step1.ANOVA [1] is generally used for significant tests of differences in means of two and more samples (i.e., factors). Among them, we choose one-way ANOVA in whether there is a significant difference between different types of glass for each of its chemical components.

Step 2. Mathematical model

In equation (6), β_{ij} is the matrix of the corresponding chemical components, a is the overall mean, and μ_i is the effect of the i level on the test index.

$$\begin{cases} \beta_{ij} = a + \mu_i + e_{ij}, i = 1, 2, \cdots, m, j = 1, 2, \cdots, n_i \\ e_{ij} \sim N(0, \sigma^2) \\ \sum_{i=1}^{m} n_i \mu_i = 0 \end{cases} \tag{1}$$

This leads to the results of the analysis of the normality test of the quantitative variables for each chemical component (see Appendix 9)

Step3.F test method

For this, we also need to analyze by the F-test method, as $\overline{X}_i \sim N\left(a_i, \frac{\sigma^2}{n_i}\right)$, $\overline{X} \sim N\left(a, \frac{\sigma^2}{n}\right)$,

$$S_A = \sum_{i=1}^{m} n_i \left(\overline{X}_i - \overline{X}\right)^2$$

$$E(S_A) = (m-1)\sigma^2 + \sum_{i=1}^{m} n_i (a_i - a)^2 \quad (2)$$

We then analyzed the results of the chi-squared test by SPSS as shown in Fig:

Table 1. Results of chi-square test

	Type (standard deviation)		F	P
	High potassium (n=18)	Lead barium(n=49)		
BaO	0.842	8.331	12.256	0.001***
PbO	0.514	14.947	42.508	0.000***
K2O	5.308	0.276	239.432	0.000***
SO2	0.157	3.139	4.026	0.049**
CaO	3.308	1.635	43.324	0.000***
CuO	1.492	2.47	1.605	0.21
MgO	0.712	0.63	0.825	0.367
SrO	0.044	0.264	17.131	0.000***
SiO2	14.467	18.646	1.711	0.196
SnO2	0.556	0.213	3.026	0.087*
P2O5	1.281	3.909	23.493	0.000***
Na2O	1.089	1.813	2.772	0.101
Fe2O3	1.566	0.948	7.739	0.007***
Al2O3	3.077	3.009	0.852	0.359

** At the 0.01 level (two-tailed), the correlation is significant. * At the 0.05 level (two-tailed), the correlation is significant.

It can be learned that the p-values of magnesium oxide, sodium oxide, copper oxide, tin dioxide, and aluminum oxide among them are all greater than 0.05, so they are not statistically significant, so it can be roughly stated that there are no significant differences between different types of glasses in these chemical compositions.

3.2. Logistic Regression Modeling for Dichotomous Classification

(1) According to the requirements of the question, we can establish the 0-1 variable to complete the prediction category as the dependent variable, 0 indicates that the unknown type of artifact is high potassium glass, 1 indicates that the unknown type of artifact is lead barium glass, and the chemical composition that we previously filtered out through ANOVA as the independent variable, to establish a silica, potassium oxide, calcium oxide, iron oxide, lead oxide, barium oxide, phosphorus pentoxide, and strontium oxide and sulfur dioxide in the logistic regression model. Therefore, when using the logistic regression model [2-3], the dummy variables were first processed by Excel and then logistic regression was performed by SPSS to predict whether the unknown artifacts were of type 0 or 1.

(2) Step1. Determine the probability of the two-point distribution.

$$\begin{cases} P(y=1|x) = F(x,\beta) \\ P(y=0|x) = 1 - F(x,\beta) \end{cases} \quad (3)$$

Step2. Take the connection function as $sigmoid$ function.

$$F(x,\beta) = S(x_i'\beta) = \frac{exp(x_i'\beta)}{1+exp(x_i'\beta)} \quad (4)$$

Step3. For the nonlinear model, the estimation is performed using the maximum likelihood estimation method.

$$f(y_i|x_i,\beta) = \begin{cases} S\left(x'_i\hat{\beta}\right) & , y_i = 1 \\ 1 - S\left(x'_i\hat{\beta}\right), & y_i = 0 \end{cases} \quad (5)$$

Step4. Bring the regression coefficient to. $\hat{\beta}$ $\hat{y}$

$$\hat{y_i} = P(y_i=1|x) = S\left(x'_i\hat{\beta}\right) = \frac{exp\left(x'_i\hat{\beta}\right)}{1+exp\left(x'_i\hat{\beta}\right)} \quad (6)$$

Where y is the actual value and $\hat{y}$ is the predicted result. Where if the predicted value $\hat{y} \geqslant 0.5$ is considered its prediction $y=1$, otherwise $y=0$.

After analyzing the above steps, the information of the components filtered by ANOVA and the corresponding information of the unknown artifacts are imported into SPSS for binary regression, and the regression coefficients obtained are $\hat{\beta}$ and the constants are:

Table 2. Regression coefficients corresponding to each filtered input variable

Variables to be entered	$\hat{\beta}$
Silicon dioxide (SiO2)	-0.198
Potassium oxide (K2O)	-1.131
Calcium oxide (CaO)	-0.329
Iron oxide (Fe2O3)	-0.87
Lead oxide (PbO)	1.545
Barium oxide (BaO)	1.643
Phosphorus pentoxide (P2O5)	-4.741
Strontium oxide (SrO)	35.408
Sulfur dioxide (SO2)	-4.229
Constants	2.005

Once the regression coefficients are found, they are brought into the predicted values at $\hat{y}$ to obtain the predicted values.

Then compare $\hat{y}$ with the actual value y to get the following table:

Table 3. Results of predicted artifact types

Artifact Number	$\hat{y}$	Predicted results
A1	0	High Potassium
A2	0	High Potassium
A3	1	Lead Barium
A4	0.99981	Lead Barium
A5	1	Lead Barium
A6	0	High Potassium
A7	0	High Potassium
A8	1	Lead Barium

3.3. Sensitivity Analysis of the Model

The above-mentioned values from the ANOVA screening showed that there were also some values with large differences, and we extracted the difference data and then used SPSS to get the final prediction results. I intend to extract the largest and smallest component data to see if it has changed (1 for high potassium and 0 for lead and barium):

Table 4. Results after extraction of sulfur dioxide

Artifact Number	y_hat	Predicted results
A1	0	1
A2	0	0
A3	1	0
A4	0.99988	0
A5	1	0
A6	0	1
A7	0	1
A8	1	0

From the above two figures, it can be seen that the extraction of sulfur dioxide has no effect on its results, while the extraction of silica has produced a change in its final classification, with the category of A2 changing from lead-barium to high potassium, which leads to the assumption that silica is a subset of its sensitivity factors.

Table 5. Results after extraction of silica

Artifact Number	y_hat	Predicted results
A1	0	1
A2	0	1
A3	1	0
A4	0.99997	0
A5	1	0
A6	0	0
A7	0	1
A8	1	0

4. Conclusion

This paper focuses on the application of ANOVA and dichotomous logistic regression model algorithms. In the study, we positioned the experimental object to the identification of glass artifact categories, and in the screening of its chemical composition by one-way ANOVA, it can be significantly found that some of the chemical components do not differ significantly in different glass artifact categories, and then the former screening results of its classification identification by regression model, and in order to make the identification results less error, and after the sensitivity analysis of its results It is clear that SiO2 is a subset of the sensitive factors in the classification results. In this process, the feasibility and accuracy of these modeling algorithms for such problems can be seen, and further research on such modeling algorithms can be developed and explored.

Acknowledgments

The authors gratefully acknowledge the financial support from Innovation and Entrepreneurship Training Program of Wuhan Business University (202211654165), Ministry of Education Industry-University Cooperative Education Project (220905181091456).

References

[1] Wenjia L, Xiaofeng Z, Lian Z. Business Process Clustering Method Based on k-means and Elbow Method[J]. J. Jianghan Univ, 2020, 48: 81-90.

[2] ZHENG Min, ZHANG Yuzheng, LV Haiyong, et al. Stata implementation method of dichotomous outcome clinical prediction model based on logistic regression [China Health Statistics, 2022,39(03):461-464.

[3] Lu Shan. Research on influencing factors of rural revitalization based on binary logistic regression model: From the perspective of grassroots cadre Contemporary Economy, 2022, 39 (07):106-110.