Ordinal Logistic Regression Analysis on Influencing Factors of Space Tourism Expectation Model

Wenjing Duan
School of Humanities, Social Sciences & Law, Harbin Institute of Technology, Harbin, Heilongjiang Province, 150000, China
19S016027@stu.hit.edu.cn Corresponding author’s e-mail: 15776670669@163.com

Abstract. Aiming at the problem of data validity in big data and information mining technology, this paper uses the method of ordinal logistic regression analysis based on statistics to evaluate the statistical validity of data, and analyses a case of space tourism expectation mode including 2025 valid samples. By analysing and calculating the significance level of the whole equation and the significance level of the equation goodness of fit in the case, the validity of the statistical data is evaluated, and the data regression equation is formed to support the statistical based artificial intelligence learning method.

1. Introduction
In recent years, with the continuous development of space technology, space economy and space industry have become more and more hot topics[1], followed by space tourism has gradually entered the public's vision[2]. Space tourism, as the main way for ordinary people to explore the mysteries of space, has also become one of the important ways for the economic development of adjacent space industry. Although the exploration of space tourism is only in its infancy, in order to better understand the people's expectations for space tourism mode and better serve the economic development of space tourism, there are several methods to solve this problem, such as inverse system decoupling [3], statistical analysis and time series analysis at present. This paper uses some data from the questionnaire of space tourism tourists' needs, and uses the statistical analysis method of ordered logistic regression, aiming at different gender, age, education background, education background and family monthly income. This paper studies the influence factors of people's expectation model of space tourism by those aspects, and draws a conclusion.

2. An overview of the principle of ordinal logistic regression analysis
Logistic regression analysis is a kind of generalized linear regression analysis with dependent variable as category variable. In the logistic regression model, the dependent variable is set as Y, obeying binomial distribution, the values are 0 and 1, and the independent variables are X₁, X₂, ⋯, Xₙ. At this time, the logistic regression model corresponding to the independent variable is as follows [4]:

\[
P(Y = 1) = \frac{\text{EXP}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}{1 + \text{EXP}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}
\]

From the simplification of the above formula, we can get that:
The logarithm of the above formula can be obtained:

$$\logit P(Y = 1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Same as linear regression model, $\beta_0$ is a constant term (or intercept), $\beta_i$ is the partial regression coefficient corresponding to $X_i$ ($i = 1, 2, m$). Compared with the traditional dependent variable which is a binary category variable, the logit transformation is $\pi_1$, $\pi_1 + \pi_2$, which is the cumulative probability of the ordered value level of dependent variables. Taking the dependent variables of three levels as an example, assuming that the values of dependent variables are 1, 2 and 3, and the probability of corresponding value levels is $\pi_1$, $\pi_2$, $\pi_3$, two models are fitted for M independent variables as follows[5]:

$$\text{Logit} \frac{\pi_1}{1 - \pi_1} = \text{Logit} \frac{\pi_1}{\pi_2 + \pi_3} = -\alpha_1 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

$$\text{Logit} \frac{\pi_1 + \pi_2}{\pi_3} = -\alpha_1 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

In fact, this model divides the dependent variables into two grades according to different value levels, and establishes a binary logistic regression model for the two grades. Similar to binary logistic regression, regression coefficient $\beta_m$ indicates that when other independent variables remain unchanged, a certain independent variable changes one unit, and the dependent variable increases the log odds ratio of one or more grades.

The dependent variable space tourism expectation model in this paper happens to be a hierarchical three category variable, so this paper chooses the ordinal logistic regression method as the main research and analysis method. Based on this, this paper mainly uses the method of Ordinal logistic regression analysis, taking gender, age, education background and family monthly income as dependent variables, and taking the expectation pattern of people for space tourism as category ordered dependent variable. According to the questionnaire, there are three types of people's expectations for space tourism, which are sightseeing mode, part participation tourism mode and whole participation tourism mode from low to high. These three modes reflect the willingness of people at different levels to participate in space tourism.

3. Ordinal logistic regression analysis

3.1. Ordinal logistic regression analysis

Because we don’t discuss the situation of cases under the age of 17, so we set all the cases under 17 years old as missing values and then conduct statistical analysis. This paper analyses the influence of gender, age, education background and family monthly income on the expectation value of public space tourism participation.

(1) Gender variable (S): when S = 1, the gender of the case is male; when S = 2, it is female.

(2) Age variable (A): when A = 2, the age of the case is between 18 and 30 years old; when A = 3, it is between 31 and 40 years old; when A = 4, it is 41 to 50 years old; when A = 5, it is 51 to 60 years old; when A = 6, it is over 60 years old.

(3) Education variable (D): D = 1 indicates that the education level of the case is junior college or below; D = 2 indicates that the education level of the case is undergraduate; D = 3 indicates that the education level of the case is master's degree; D = 4 indicates that the education level of the case is doctor's degree.
(4) Income variable (I): when I = 1, the monthly household income is less than 5000 yuan; when I = 2, it is between 5001 and 10000 yuan; when I = 3, it is between 10001 and 20000 yuan; when I = 4, it is between 20001 and 40000 yuan; when I = 5, it is between 40001 and 60000; when I = 6, it is between 60001 and 80000 yuan. When I = 7, the family monthly income is between 80001 yuan and 100000 yuan; when I = 8, it means that the family monthly income is more than 100000 yuan.

(5) The variable of expectation model (M): according to the degree of participation from low to high, M = 1 indicates that the expected mode is sightseeing mode; M = 2 indicates that the expectation mode is partial participation tourism mode; when M = 3, it indicates that the expectation mode is full participation tourism mode = 3.

The above variables and the number of investigation cases corresponding to each variable are plotted as follows:

| Case summary | Number of cases /N | Marginal percentage |
|--------------|--------------------|---------------------|
| Expected space tourism mode (M) | | |
| 1 | 473 | 23.4% |
| 2 | 1202 | 59.4% |
| 3 | 350 | 17.3% |
| Sex(S) | | |
| 1 | 1020 | 50.4% |
| 2 | 1005 | 49.6% |
| Diploma (D) | | |
| 1 | 297 | 14.7% |
| 2 | 1247 | 61.6% |
| 3 | 339 | 16.7% |
| 4 | 142 | 7.0% |
| Income of a family (I) | | |
| 1 | 250 | 12.3% |
| 2 | 535 | 26.4% |
| 3 | 621 | 30.7% |
| 4 | 427 | 21.1% |
| 5 | 93 | 4.6% |
| 6 | 25 | 1.2% |
| 7 | 27 | 1.3% |
| 8 | 47 | 2.3% |
| Age (A) | | |
| 1 | 13 | 0.6% |
| 2 | 1054 | 52.0% |
| 3 | 589 | 29.1% |
| 4 | 304 | 15.0% |
| 5 | 53 | 2.6% |
| 6 | 12 | 0.6% |
| Total | 2025 | 100% |

A total of 2025 effective cases were obtained, and the above table is the case value.

3.2. Model fitting information
After fitting the above information, the degree of freedom, significance level and other statistical information of the model were calculated. The fitting information is as follows:
Table 2. Overall fitting information of the model

| Fitting information | -2log(likelihood) | $\chi^2$ | degree of freedom(df) | significance level |
|---------------------|-------------------|---------|------------------------|--------------------|
| Data                | 947.811           | 118.702 | 16                     | 0.000              |

The above is the fitting information of the whole model. The purpose of fitting is to test the regression of the equation as a whole. By calculating the significance level of the fitting equation, we can judge whether the whole logistic regression equation has statistical significance and whether it can be used for information mining. According to the model fitting calculation results, the chi square statistic of the model is 118.702, and the significance level of the logistic regression equation is far less than 0.05, which indicates that at least one independent variable is statistically significant. Therefore, the logistic regression equation is meaningful and can be used for subsequent statistical analysis and information mining.

3.3. Goodness of fit analysis of logistic regression equation

The goodness of fit of logistic regression equation of the model was analysed. Pearson correlation coefficient and goodness of fit deviation of logistic regression equation were selected to calculate the significance level of goodness of fit. The calculation results are shown in the table below:

Table 3. Model goodness of fit information

| Model | $\chi^2$ | degree of freedom(df) | significance level of goodness of fit |
|-------|---------|------------------------|--------------------------------------|
| Pearson | 565.014 | 368                    | 0.000                                |
| Bias   | 473.275 | 368                    | 0.000                                |

The table above is a test of the goodness of fit of the equation. The goodness of fit of the equation is used to judge the explanatory power of the independent variable of the equation to the dependent variable. Through the table data, it can be found that the p value of the fitting degree is far less than 0.05, and the goodness of fit of the equation is generally significant.

3.4. Interpretation and analysis of ordinal logistic regression equation

Table 4. Parameter estimation of regression coefficient of equation

| Variable | Estimation | Standard error | Wald | df | Significance | 95%Confidence interval |
|----------|------------|----------------|------|----|--------------|------------------------|
|          |            |                |      |    |              | Lower limit           |
|          |            |                |      |    |              | Upper limit           |
| M=2      | -0.431     | 0.671          | 0.413| 1  | 0.520        | -1.747                |
|          |            |                |      |    |              | 0.884                 |
| M=3      | 2.443      | 0.673          | 13.191| 1 | 0.000        | 1.125                |
|          |            |                |      |    |              | 3.762                 |
| S=1      | -0.032     | 0.089          | 0.129| 1  | 0.719        | -0.206               |
|          |            |                |      |    |              | 0.142                 |
| S=2      | 0          | 0.0            | 0.0  | 0  | 0.0          | 0.0                   |
|          |            |                |      |    |              | 0.0                   |
| D=1      | -0.265     | 0.206          | 1.656| 1 | 0.198        | -0.669               |
|          |            |                |      |    |              | 0.139                 |
| D=2      | -0.218     | 0.179          | 1.481| 1 | 0.224        | -0.569               |
|          |            |                |      |    |              | 0.133                 |
| D=3      | 0.048      | 0.201          | 0.058| 1 | 0.810        | -0.346               |
|          |            |                |      |    |              | 0.443                 |
| D=4      | 0          | 0.0            | 0.0  | 0  | 0.0          | 0.0                   |
|          |            |                |      |    |              | 0.0                   |
| I=1      | -0.534     | 0.319          | 2.805| 1 | 0.094        | -1.160               |
|          |            |                |      |    |              | 0.091                 |
| I=2      | -0.460     | 0.304          | 2.288| 1 | 0.130        | -1.056               |
|          |            |                |      |    |              | 0.136                 |
| I=3      | -0.399     | 0.301          | 1.755| 1 | 0.185        | -0.989               |
|          |            |                |      |    |              | 0.191                 |
| I=4      | -0.387     | 0.306          | 1.602| 1 | 0.206        | -0.986               |
|          |            |                |      |    |              | 0.212                 |
| A   | 1.354 0.798 2.882 1 0.090 -0.209 2.918 |
|-----|-----------------------------------------|
| A=2 | 1.689 0.580 8.483 1 0.004 0.552 2.825 |
| A=3 | 1.363 0.583 5.465 1 0.019 0.220 2.507 |
| A=4 | 0.710 0.586 1.466 1 0.226 -0.439 1.858 |
| A=5 | -0.483 0.644 0.563 1 0.453 -1.745 0.779 |
| A=6 | 0 0 0 0 0 0 0 |

Suppose that the participation mode of tourists' expectation for space tourism is $y$, and the significance in the table indicates the significance level of each independent variable in the regression equation as a whole. The estimated column is the parameter value of the equation. Therefore, the overall equation of the regression equation can be obtained as follows:

$$Y = 0.32S(1) - 0.265D(1) - 0.218D(2) + 0.048D(3) - 0.534I(1) - 0.46I(2) - 0.399I(3) - 0.387I(4) - 0.408I(5) + 0.149I(6) + 0.227I(7) + 1.354A(1) + 1.689A(2) + 1.363A(3) + 0.710A(4) - 0.483A(5)$$

### 4. Conclusion

Through the fitting analysis of the independent variables and dependent variables in this paper, the overall significance of the equation and the goodness of fit are significant. The statistical analysis of this paper is of statistical significance and can be used as the basis for data mining processing. From the equation, we can get that:

1. Gender has little influence on the expected participation mode of space tourism;
2. Educational background has a certain influence on the expected participation mode of space tourism. Through the equation, we can find that the higher the educational background, the more the tourists expect to participate in space tourism;
3. As for the independent variable of family income, the higher the family income is, the higher the expectation for the participation mode of space tourism is. After fitting the model, the Wald test was carried out on the significance level of the coefficients. In the test, the significance of the coefficients was basically less than 0.05, so it could be judged that the coefficients were significant in the equation;
4. For age, there is a positive correlation distribution. When the age is below 60 years old, the expectation value of space tourism participation increases with the increase of age. However, after the age of 60 years old, the expectation value of space tourism participation decreases due to age growth and physical reasons.

So we can design space tourism product model to meet the needs of different levels of tourists through the date.

### References

[1] Billings, L. (2006). Exploration for the masses? or joyrides for the ultra-rich? prospects for space tourism. Space Policy, 22(3), 162-164.
[2] David Bell. (2009). Space travel and culture: from apollo to space tourism. Journal of Clinical Endocrinology & Metabolism, 14(6), 672-3.
[3] Peng Wu, Rongjun Mu, Xin Zhang, & Yanpeng Deng. (2018). Multiple estimation channel decoupling and optimization method based on inverse system. Young Scientists Forum. Young Scientists Forum 2017.
[4] Kianifard, F, & Kleinbaum, D. G. (2010). Logistic regression: a self-learning text. Technometrics, 37(1), 116-117.

[5] Brant, R., (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. Biometrics, 46(4), 1171-1178.