ISSN: 2664-9640

# Logistic regression model based on LPM

Haisong Xu

Department of Electric Power Engineering, North China Electric Power University, Baoding 071000 , China

220191030221@ncepu.edu.cn

## Abstract

**Established multiple regression model based on the factors affecting the survival of wasps and expansion model based on regional weight and wasp migration radius ; collect the data and sort out the significant traits of wasps and other wasps, establish a logistic regression model based on LPM.**

## Keywords

**Multivariate linear regression; based on LPM logical regression; step-by-step screening ; space expansion.**

## 1. Introduction

In problem 1, in order to predict the spatial distribution of wasps over time, I do simple data processing, and then find that there is a correlation between environmental variables and the number of witnesses. Then, I established a multiple regression model based on the factors affecting the survival of wasps (the standardized regression coefficients of the factors affecting the distribution of wasps) and an expansion model based on the regional weight and the migration radius of wasps (defining the trend degree score, and calculating the proportion of wasps population density and the expected number of wasps in the next eight years). Finally, I test the robustness and accuracy of the model (test statistics, heteroscedasticity, multicollinearity, discussion of goodness of fit);

In problem 2, in order to solve the problem of the possibility of false prediction, I first collected data and sorted out the significant characters (size, yellow head, black chest, only striped abdomen, yellow black abdomen, hairy or not, yellow tail) that distinguish wasps and other bees, and then screened out the image samples with high definition and strong judgment So I established a logistic regression model based on LPM, and used the results to judge whether it was wasp or not.I take the prediction result $y_i$ as the probability that a sample is a wasp , if $y_i \geq 0.5$,The sample is believed to be a $y_i$=1（a wasp）； if $y_i < 0.5$,The sample is believed to be a $y_i$=0（not a wasp）； Finally, the accuracy of the model was tested (training group and cross validation with test group).

Text:the solution of problem one: establish the spatial expansion model of wasps with time

Data preprocessing

I regard the map of Washington state as a rectangle, and divide it into 16 small areas evenly as shown in Figure 1, and number them by 1 ~ 16. Then I use ArcGIS to process the longitude and latitude corresponding to the eyewitness events, so as to get the number of eyewitness events in each small area, and regard all the number of eyewitness events as the number of wasps.Use the website to query the average temperature, average precipitation and altitude (as the average temperature, average precipitation and air pressure) [1] of the main cities in each small area, record the number of main cities in each small area, and make a visual list of the data, as shown in the Figure 2.

ISSN: 2664-9640

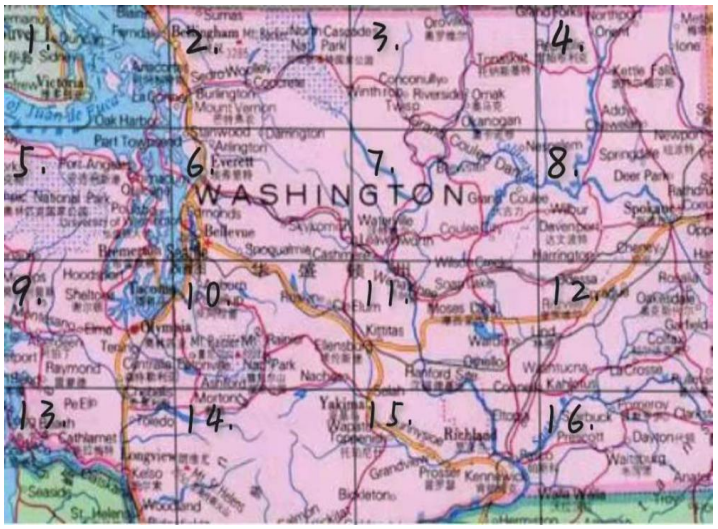


Figure 1 Data Description

| Serial Number | Mean Value of Hornet | Average Temperature | Average Annual Precipitation( | Average Number of Major | Pressure(mbar) |
|---|---|---|---|---|---|
| 1 | 253 | 9 | 531 | 3 | 1018 |
| 2 | 642 | 10 | 698.2 | 2 | 1018 |
| 3 | 31 | 6 | 531 | 1 | 1018 |
| 4 | 23 | 4 | 177.4 | 3 | 1017 |
| 5 | 271 | 9 | 391.9 | 9 | 1018 |
| 6 | 2039 | 14 | 848.7 | 2 | 1018 |
| 7 | 343 | 10 | 262.8 | 2 | 1017 |
| 8 | 228 | 10 | 148.5 | 1 | 1017 |
| 9 | 315 | 10 | 585.6 | 5 | 1018 |
| 10 | 340 | 10 | 505.6 | 1 | 1018 |
| 11 | 1411 | 12 | 703.6 | 4 | 1016 |
| 12 | 74 | 8 | 79.8 | 2 | 1017 |
| 13 | 84 | 9 | 77.6 | 3 | 1018 |
| 14 | 141 | 11 | 703.6 | 2 | 1018 |
| 15 | 100 | 9 | 77.5 | 5 | 1016 |
| 16 | 10 | 4 | 167.1 | 1 | 1017 |

Figure 2 Data visualization list

Establish a multiple regression model based on the factors affecting the survival of wasps

I regard the average temperature, average precipitation, atmospheric pressure and the number of main cities in a certain area as the core factors that affect the distribution of wasps (that is, the core factors that affect the distribution of wasps)[2], so I take the number of wasps in a certain area as the dependent variable and the influencing factors as the independent variable, and establish a multiple regression model based on the factors that affect the survival of wasps[3].

$$number_i = \alpha + \beta_1 \times X_1 + \beta_2 \times X_2 + \beta_3 \times X_3 + \beta_4 \times X_4 + \varepsilon_i \quad (1)$$

The $number_i$ here represents the number of bees in the area, $X_1$、$X_2$、$X_3$、$X_4$ represent the average temperature, average precipitation, number of main cities and air pressure of a certain area.After adding the control variables, I will use Stata to estimate all the regression coefficients, and use F statistics to test whether the regression coefficients $\beta_1=\beta_2=\beta_3=\beta_4=0$ are joint significant, and find out the environmental variables that have a significant relationship with the number of bees in a certain area[4],The experimental data is shown in figure 3.

ISSN: 2664-9640

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| AverageTem~e | 16 | 9.0625 | 2.619637 | 4 | 14 |
| AverageAnn~m | 16 | 405.6188 | 264.629 | 77.5 | 848.7 |
| AverageNum~s | 16 | 2.875 | 2.093641 | 1 | 9 |
| Pressurembar | 16 | 1017.438 | .7274384 | 1016 | 1018 |
| number | 16 | 8.5 | 4.760952 | 1 | 16 |

Figure 3 Data preprocessing of Stata regression analysis

Model conclusion: according to F-statistic test regression and t-statistic test regression, only the average temperature and precipitation can be significantly different from 0 at 90% confidence level, that is, I think that the average temperature and precipitation have a significant correlation with the number of bees. At the same time, in order to more accurately study the important factors that affect the evaluation (excluding the influence of dimension), I consider using standardized regression coefficient for comparison.

To standardize the data is to subtract its mean from the original data, and then divide it by the standard deviation of the variable to calculate the new variable value. The regression equation formed by the new variable is called the standardized regression equation, and the standardized regression coefficient can be obtained after regression.

Finally, I get the standardized regression coefficients of average temperature and precipitation are 0.4523635 and 0.574374295.

An expansion model based on regional weight and migration radius of wasps is established

First, I define a new variable according to two standardized regression coefficients, which I call trend score M. The tendency score m refers to the invasion tendency of wasps to a certain area, that is, the higher the tendency score is, the more likely wasps are to invade the area.

$$M = 0.4523635 \times X_1 + 0.574374295 \times X_2 \quad (2)$$

I calculate the trend score of each region in figure 4, then standardize it, and make the standardized probability visual map in figure 5.

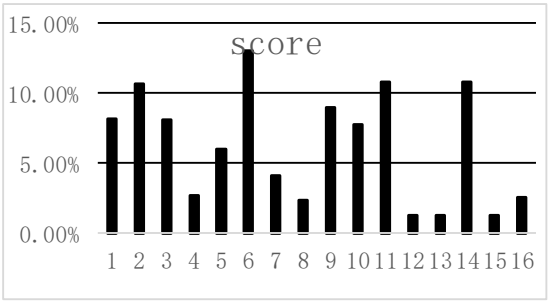


Figure 4 the trend score of each region

| | | | |
|---|---|---|---|
| 8.15% | 10.69% | 8.11% | 2.73% |
| 6.04% | 13.01% | 4.10% | 2.37% |
| 8.98% | 7.77% | 10.79% | 1.31% |
| 1.29% | 10.78% | 1.29% | 2.58% |

Figure 5 Standardized probability visual graph

According to the area of Washington state, the area is 176676 square kilometers. Due to the rectangular area formula, I approximately think that the edge length of the 16 small areas

ISSN: 2664-9640

divided as shown in Figure 6 is 120 kilometers and 90 kilometers. For the convenience of calculation, I assume that all the confirmed wasps are concentrated at the upper critical points of zone 1 and zone 2, and all 34 eyewitness reports are concentrated in zone 2. I let the wasps expand outward from the area (approximately as a point), because the nesting distance of a new queen is estimated to be 30 km, so it can be approximately considered that the expansion range of the wasp in the first year is within the semicircle with the critical point as the center, and the radius of the circle expands outward by 30 km every year in the following years. It is assumed that the number of wasps in the circular area will not decrease every year.

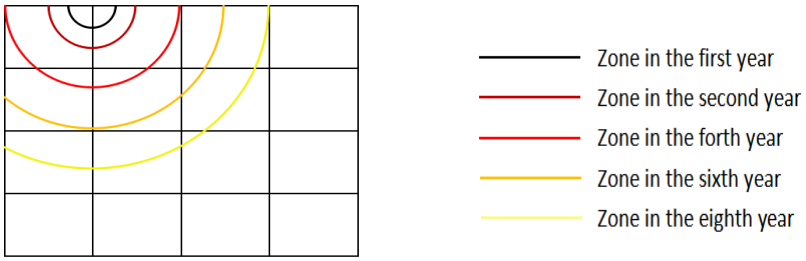


Figure 6 Migration model of Asian giant hornet

Let the intermediate variable as $m_i = w_i \times s_i$ ; $w_i$ be the weight of area I (the score of trend degree); $S_i$ be the expanded area of wasps in area I in a certain year; the population density ratio in I region in a certain year is $p_i = \frac{m_i}{\sum_{i=1}^{16} m_i}$ ; Assuming that the second area always corresponds to 34 wasp nests, the expected numbers of each area in the second, fourth, sixth and eighth years are calculated respectively $N_i = \frac{34P_i}{P_2}$ （i=1, 3~16）, so I get the figure 7.

| | the second year | the forth year | the sixth year | the eighth year |
|---|---|---|---|---|
| N1 | 26 | 26 | 26 | 26 |
| N2 | 34 | 34 | 34 | 34 |
| N3 | 0 | 0 | 11 | 25 |
| N4 | 0 | 0 | 0 | 0 |
| N5 | 0 | 3 | 15 | 19 |
| N6 | 0 | 7 | 32 | 41 |
| N7 | 0 | 0 | 2 | 9 |
| N8 | 0 | 0 | 0 | 0 |
| N9 | 0 | 0 | 0 | 16 |
| N10 | 0 | 0 | 0 | 14 |
| N11 | 0 | 0 | 0 | 0 |
| N12 | 0 | 0 | 0 | 0 |
| N13 | 0 | 0 | 0 | 0 |
| N14 | 0 | 0 | 0 | 0 |
| N15 | 0 | 0 | 0 | 0 |
| N16 | 0 | 0 | 0 | 0 |

Figure 7 Prediction model of hornet population in different regions

The relative size of the data $P_i$ in the table reflects the relative size of the number of wasps in each region. Based on this, Figure 8 was constructed, and the wasp numbers in the second, fourth, sixth and eighth years Ire used as the prediction results.

ISSN: 2664-9640

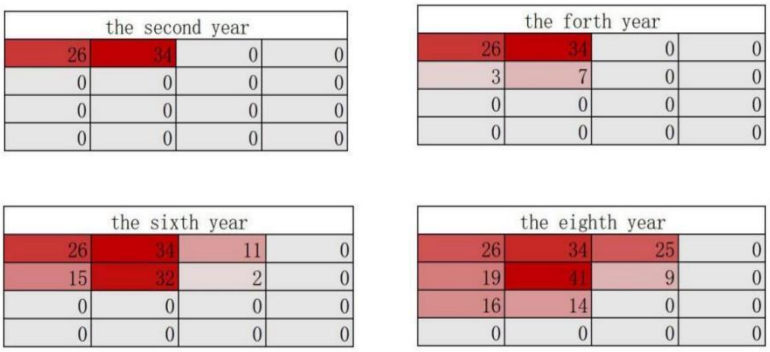


Figure 8 Number of Asian giant hornet in different years

According to the analysis of the graph, the dark or light color of the region represents the relative number of wasp population, and the number of the darkest region in the four graphs above is 41. In the second year, the wasp population only distributed in the first and second regions; in the fourth year, the wasp population expanded to the fifth and sixth regions, and most of them still distributed in the first and second regions; similarly, in the sixth year, the wasp population expanded to the third and seventh regions, and the distribution center shifted to the second and sixth regions; in the eighth year, the wasp population expanded to the ninth and tenth regions, and the distribution center shifted from the second region to the sixth region Domain.

Based on this model, I can approximately predict the area of wasp expansion in Washington state and the proportion of wasp density in each area.

3. Solution to Problem 2: Create, analyze and discuss a model to predict the possibility of misclassification

Collect data and sort out the significant traits of wasps and other wasps

I selected seven characteristics that can distinguish wasps from other wasps according to the data and image resources: size, yellow head, black chest, only striped abdomen, yellow black abdomen, hair and yellow tail.Then I selected 17 reports with high definition and strong judgment from the selected negative reports. Because the number of positive reports is too small, I use all 13 positive reports to form a set of 30 reports that can be clearly judged by the pictures. Then I transformed the report set into the wasp character matching table based on the selected seven characters and 30 image data.

Establish a logistic regression model based on LPM

Data processing instructions:

According to the two-point distribution, I use 1 to represent the samples that have been identified as wasps, and 0 to represent the samples that are not wasps. Then I will confirm that the wasp character is represented by 1, not the wasp character is represented by 0, and get 30 groups of data. Here I use SPSS to calculate the logistic regression model. Because whether the prediction result is a wasp is a qualitative variable, I create a dummy variable to represent the prediction result. If the prediction result is closer to 1, it means that the sample is a wasp, and closer to 0, it means that it is not a wasp.

Linear probability model (LPM)[5][6]:

$$y_i = \beta_0 + \beta_1 X_{1+} \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \mu_0 \quad (4)$$

Because of the endogeneity of LPM model (the prediction value is not accurate) and the prediction does not conform to the actual situation (the prediction probability may be greater than 1 or less than 0), So I use the join function $F(x, \beta) = S(\acute{x}_1\beta) = \frac{\exp(\acute{x}_1\beta)}{1+\exp(\acute{x}_1\beta)}$ here to ensure that the range of prediction results is between 0-1.

ISSN: 2664-9640

Logistic regression model：

Here, I set Y (dependent variable) for wasp or not, and set the other seven traits as x1, X2, X3, x4, X5, X6, X7 (independent variable) to establish the following model.

Because I think this model is nonlinear, I use the maximum likelihood estimation method to estimate $\beta_i$.

$$y_i = p\left(y_i = 1|x\right) = S\left(x_i'\beta\right) = \frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)} = \frac{e^{\beta_0+\beta_1X_{1+}\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7}}{1+e^{\beta_0+\beta_1X_{1+}\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7}} \quad (5)$$

Then I use SPSS to import data to solve the logistic regression model.

The results Ire used to judge whether it was wasp or not：

I take the prediction result $y_i$ as the probability that a sample is a wasp,if $y_i \geq 0.5$,The sample is believed to be a $y_i$=1（a wasp）； if $y_i < 0.5$, The sample is believed to be a $y_i$=0（not a wasp）.Data processing is shown in figure 9.

| Unweighted cases[a] | | cases | percentage |
|---|---|---|---|
| Selected cases | included cases | 24 | 80.0 |
| | missing cases | 6 | 20.0 |
| | total | 30 | 100.0 |
| Unselected cases | | 0 | .0 |
| total | | 30 | 100.0 |

Figure 9 Summary of case handling

## 2. Conclusion

Because the existence of wasps will greatly damage the local ecological balance and have a greater impact on people's life and production, it will cause anxiety of the government and local people. Therefore, in-depth analysis of the data provided by the public report, research efficient and accurate strategy to give priority to the suitable public report for further investigation is very meaningful.

In this issue, I have done in-depth data cleaning for the data set provided by the public report. After that, I visualized the processed data from multiple dimensions to explore the possible correlation between the data from various angles. Finally, I analyze the environmental variables, the number of reports and image features in the time and space dimensions, and comprehensively use multiple regression, weight based spatial expansion, logistic regression, distribution screening, Gaussian fitting and other models to further explain the data provided by the public report, and dig out the unique and important opinions of the public report, so as to provide reference for the Washington state government Provide effective suggestions.

## References

[1] Qiu Yong, Guo yunjiao, Yang Xinzhou, Yang Qi, Li Zhaoyun. Key techniques for artificial breeding of wasps [J]. Livestock and poultry industry, 2020,31 (12): 51-52.

[2] Wang Lin, Miao Gang He. Establishment and analysis of multiple linear regression model for the number of firefighters killed in duty [J]. Fire protection today, 2020,5 (12): 50-51.

ISSN: 2664-9640

[3] Williams Charles Gbenga,Ojuri Oluwapelumi O.. Predictive modelling of soils' hydraulic conductivity using artificial neural network and multiple linear regression[J]. SN Applied Sciences,2021,3(2).

[4] Valentini Marlon,dos Santos Gabriel Borges,Muller Vieira Bruno. Multiple linear regression analysis (MLR) applied for modeling a new WQI equation for monitoring the water quality of Mirim Lagoon, in the state of Rio Grande do Sul—Brazil[J]. SN Applied Sciences,2021,3(1).

[5] Li Ying,Wang Liguo. RNA Coding Potential Prediction Using Alignment-Free Logistic Regression Model.[J]. Methods in molecular biology (Clifton, N.J.),2021,2254.

[6] JIANG Fengyang,GUAN Zhidong,LI Zengshan,WANG Xiaodong. A method of predicting visual detectability of low-velocity impact damage in composite structures based on logistic regression model[J]. Chinese Journal of Aeronautics,2021,34(1).