University of California, Davis

Statistics

House Sparrow Survival Classification

Author: Yu Zhu, Yanan Gao

Student ID: 914480813, 913556858

House Sparrow Survival Classification

Yu Zhu, Yanan Gao

March 21, 2018

Abstract

After a severe winter storm in 1898, some house sparrows found perished while others survived. In order to analyze the relation between their survival and 10 physical characteristics, we perform logistic regression on the standardized data. Likelihood ratio test and Wald test are used to select first order model, then significant variables selected are used to fit a second order model. Finally, the final model has been reached using stepwise selection (AIC) of second order full model. Pearson Residual and its loess plot suggests that the model assumption of equal variance holds.

1 Introduction

The house sparrow (Passer domesticus) is a bird of the sparrow family Passeridae, found in most parts of the world. Though it is widespread and abundant, its numbers have declined in some areas[1]. Therefore, the understanding of impacts of the house sparrow survival is of great importance. The investigation was conducted with the house sparrows found on the ground after a severe winter storm in 1898. Some sparrows survived and some perished. The ecologist recorded 10 physical characteristics: age, total length, alar extent, weight, length of beak and head, length of humerus, length of femur, length of tibio-tarsus, width of skull and length of keel of sternum. Most of the variables are describe the length of the sparrow. The whole dataset has 11 variables and 87 observations.

In this report, we want to find out the most important variables deciding the survival of a sparrow with logistic linear regression model.

2 Preliminary Study and Project Plan

2.1 Data summary and overview

The data summary is shown in Appendix. We can observe that there is no missing value. Firstly, we make the transformation of the variable: 'STATUS', substituting 'Survival' for '1' and 'Perished' for '0'(In binary logistic regression, the outcome

is usually coded as '0' or '1', as this leads to the most straightforward interpretation[2]). Secondly, by checking the type of the variables, we find that the variable 'STATUS is 'character, others are all 'numeric'. However, the variable 'STATUS' and independent variable 'AG' should be qualitative, so we factorize them at the beginning. Thirdly, for the two qualitative variables, pie charts (figure 1) shows that 'AG' has two levels, the percentage of 'adult' is 68%, which is twice as much as 'juvenile'. 'STATUS' has two levels as well, the the percentage of 'survived' is 59%, which is only a little bit more than 'perished'.

Judging by the titles of these variables literally, we speculate that the 'TT' (total length) has some linear relation with the other length variable (e.g. 'BH', 'HL', 'FL'...).

2.2 Remedy Multi-collinearity and Outliers

Before fitting model, the independent variables are first checked. As observed, since the independent variables have different units and magnitude, standardization may reduce the numerical instability(STA 206). Under this circumstance, we standardize all the continuous variables.

For the outliers, by checking the boxplots (figure 2), seven outliers whose standard deviations are larger than 3, are spotted. They correspond to the 9th, 14th, 56th, 58th, 65th, 75th and 81th observations. However, considering that the deviations are not so severe and the sample is quite small, we choose not to delete any observation.

Furthermore, we want to find out the relations among these variables. From pairwise scatter plot (figure 3): 'HL', 'FL' and 'TT' are showing obviously linear pattern, indicating the existence of colinearity. Besides, many of the remaining independent variables are also somewhat correlated.

From correlation matrix between all quantitative variables (figure 4), we can see that 'FL' and 'TT' are highly correlated with other variables, which is consistent with our previous speculation. Further analysis is needed to decide whether any variables can be dropped.

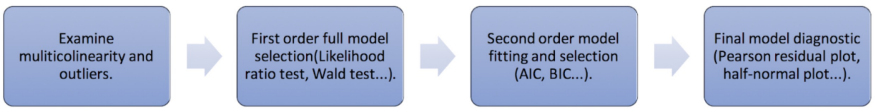

Figure 1: Work Flow

2.3 Initial Model Fitting

Since 'STATUS' is a binary dependent variable, we prefer to fit a logistic linear model. If we directly use 'STATUS' as the response variable and try to fit a line, it wouldnt direct to good result: (1)the error terms are heteroskedastic; (2)residuals

can not be normally distributed because P takes on only two values, violating the regression assumption; (3)The predicted probabilities can be greater than '1' or less than '0'[3].

Since the logistic linear model can solve these problems and easier to work with, we decide to use it to fit the data. We initially fit a first-order model with all variables and use it find out whether transformations are needed. The form of the model is:

$$Log(\frac{\pi_j}{1-\pi_j}) = \beta_0+\beta_1 AG+\beta_2 TL+\beta_3 AE+\beta_4 WT+\beta_5 BH+\beta_6 HL+\beta_7 FL+\beta_8 TT+\beta_9 SK+\beta_{10} KL$$

First, to investigate if fitting a logistic linear model is appropriate, we use Likelihood ratio test(Deviance goodness-of-fit test).

$$H0 : \pi'_j = \beta_0+\beta_1 AG+\beta_2 TL+\beta_3 AE+\beta_4 WT+\beta_5 BH+\beta_6 HL+\beta_7 FL+\beta_8 TT+\beta_9 SK+\beta_{10} KL$$

$$H1 : \pi'_j \text{do not line on straight line}$$

Since G^2 equals to 65.838, degree of freedom is 76, we compare it with χ^2 (0.95,76), which equals to 97.35097. Clearly, $G^2 < 97.35097$, so we cannot reject H0, which means fitting a logistic linear model is reasonable.

Besides, from the summary of the model (table 1), only 'TL' and 'WT' are significant at 5% level; according to the plot of Pearson residual and its loess (figure 5), it shows a constant variance pattern; from the half normal plot(figure 6), there is no severe outlier as well.

3 Model Selection and Diagnostic

3.1 Hypothesis Tests to Drop Variables

Previously, we have speculated that there are some relations between all the length related variables (i.e. some of them might be correlated), in addition, from the correlation matrix, we have also found out that 'TT' and 'FL' are highly correlated. Therefore, we first checked if these two variables can be dropped together from our first order full model:

$$H0 : \beta FL = \beta TT = 0$$

$$H1 : \text{Not all of the above coefficients are zero.}$$

From likelihood ratio test, G^2 is 0.607, which is less than χ^2 (0.95; 2) at 5% level, indicating 'FL' and 'TT' can be dropped, we name the reduced model 'fit 1'. Seen from the summary of 'fit 1', judging from the p values, many variables are with pvalues larger than 0.05, indicating these variables might be dropped as well. Therefore, we perform Wald test starting from the variable with the largest

p values, and proceed in a descending order until all variables in our model are significant. All the hypothesis tests and results are shown in the table below

Hypothesis test	Z-value	P-value
H_0: $\beta_{AG} = 0$. H_1: $\beta_{AG} \neq 0$.	0.154	0.878
H_0: $\beta_{AE} = 0$. H_1: $\beta_{AE} \neq 0$.	0.764	0.445
H_0: $\beta_{SK} = 0$. H_1: $\beta_{SK} \neq 0$.	1.050	0.294
H_0: $\beta_{BH} = 0$. H_1: $\beta_{BH} \neq 0$.	1.254	0.210

Z values are smaller than $\Phi(0.975)$, indicating the variables of 'AG', 'AE', 'SK' and 'BH' can be dropped. Therefore, we have reached to a model 'fit 2' with only 'TL, 'WT', 'HL' and 'KL' as independent variables.

3.2 Second Order Model Selection and Diagnostic

We want to use AIC criterion to help us select the final model. We use all the independent variables in model(fit 2) to fit a second-order logistic linear model as the full model.

Meanwhile, we fit a null model with no X variables. AIC of the null model is 120.01, full model is 90.11648. After the step AIC procedure, we obtain the final model with the smallest AIC: 79.771.

As for the BIC criterion, the final model is still the same. BIC of the null model is 122.4743, full model is 127.1051, and the final model has the smallest BIC: 92.10031. So our final model is :

$$Log(\frac{\pi_j}{1-\pi_j}) = \beta_0 + \beta_1 TL + \beta_2 HL + \beta_3 WT + \beta_4 KL$$

According to the plot of Pearson residual and its loess (figure 7), it also shows a constant variance pattern. So we decide to choose this model as our final model in this project.

4 Conclusion

4.1 Summary of Findings

From the summary table of final model (table 2), we find out that all the independent variables are first order. Judging from the p values, all the variables are with p values smaller than 0.05. The variables 'TL' and 'WT' both have negative effects to the survival of a sparrow, while the variables 'HL' and 'KL' both have positive effects.

Hence, we can interpret the final model: for every standard deviation unit increase in total length, the adds of a sparrow which is survived is estimated to decrease by 86.8%; for every standard deviation unit increase in the length

of humerus, the adds of a sparrow which is survived is estimated to increase by 398.56%; for every standard deviation unit increase in weight, the adds of a sparrow which is survived is estimated to decrease by 65.69%; for every standard deviation unit increase in the length of keel of sternum, the adds of a sparrow which is survived is estimated to increase by 150%.

4.2 Conclusions of the Study

After the model establishment, we can conclude that the different physical characteristics of sparrows result in different influences on the probability of survival. Among the 10 different physical characteristics, the total length, the length of humerus, weight and the length of keel of sternum have more significant effects on the survival probability, and the length of humerus causes the most significant influences.

To be more specific, if a sparrow has larger length of humerus and keel of sternum, and smaller total length and number of weight, it will have much larger survival probability. In opposite, if the sparrow with smaller length of humerus and keel of sternum, and larger total length and number of weight, it will be more likely to perish in the winter storm.

4.3 Further Improvements

Considering the high multicolinearity of this data, a more efficient way of PCA (Principle Component Analysis) can be adopted here to transform the standardized data Z. The general transforming way is:

$$Y = PZ$$

Where P is the eigen vector matrix of the sample correlation matrix of X (also the sample covariance matrix of Z). Y is then further considered as independent variables. The nice property about Y is the correlation between different y is zero. However the interpretation of the coefficient corresponding to each y is quite different, and is beyond the scope of this report.

5 References

[1] Wikipedia, House Sparrow
[2] Wikipedia, Logistic Regression
[3] Hosmer, David W.; Lemeshow, Stanley (2000).
Applied Logistic Regression (2nd ed.). Wiley. ISBN 0-471-35632-8.
[4] STA 206, 207