# House Sparrow Survival Prediction Using a Logistic Binary Regression

Fanchao(Franco) Meng 23370209

16 October, 2022

## Executive Summary

This report is to investigate the relationship between house sparrow survival rate and their morphological measurements. Data on 163 house sparrow is collected by Anatomical Laboratory of Brown University after an uncommonly severe winter storm. A logistic binary regression was fitted to the data. The model indicates that overall, sparrows with shorter total length and lighter body weight, but longer sternum and humerus have a higher chance of survival. In addition being male also increases the chance of survival.

## 1 Introduction

In order to better understand the natural selection process, and to investigate whether the survival of the house sparrows, is related to certain physical characteristics,which enable them to withstand the intensity of selective elimination. Pugesek, Bruce H. (1996) used structural equation modelling techniques, found out that survival increased significantly with increasing general size. Wing length, which was independent of its relationship to the general size factor, was also significantly in their model, where higher survival was found among birds with shorter wings. Male also had higher survival rate comparing with female. Brown, Charles R. (1998) also thoroughly investigated the similar natural selection process by using cliff swallows survival data from different years. The research found that survivors had lower wing and outer tail asymmetry, low asymmetry in wings and tail made foraging more efficient and less costly. Mortality patterns did not differ by sex, but older birds suffered heavier mortality.

Our analysis will be based on Bumpus data to investigate the relationship between survival rate and the morphological measurements. The data contains 136 house sparrow information recorded by anatomical laboratory of brown university, after an uncommonly severe winter storm in 1989. Some of these birds had survived and some had perished.

## 2 Methodology

We will use R statistical tool, firstly to explore the data by using numerical summaries, along with an correlation heat map. A logistic regression for binary count model will be fit to the data, the log survival odds ratio will be our response variable. The interactions will be explored and discussed, on whether any interactions should be included in our final model. The final model will be interpreted to explain which morphological attributes of house sparrow have significant influences on survival rate.

# 3 Results

## Summary Table

A summary of the variables in the data set has been given in the table below. The footnote has demonstrated the meaning of "zero" and "one" in the summary. Due to the fact that there are nearly double of the male records than female records, we've decided to use male (0) as our reference level. The perished sparrows are noted with (0), in order to better predicted the survival rate. Overall the data is satisfactory. All the numeric measurement seems normal distributed, no obvious skewness in the data.

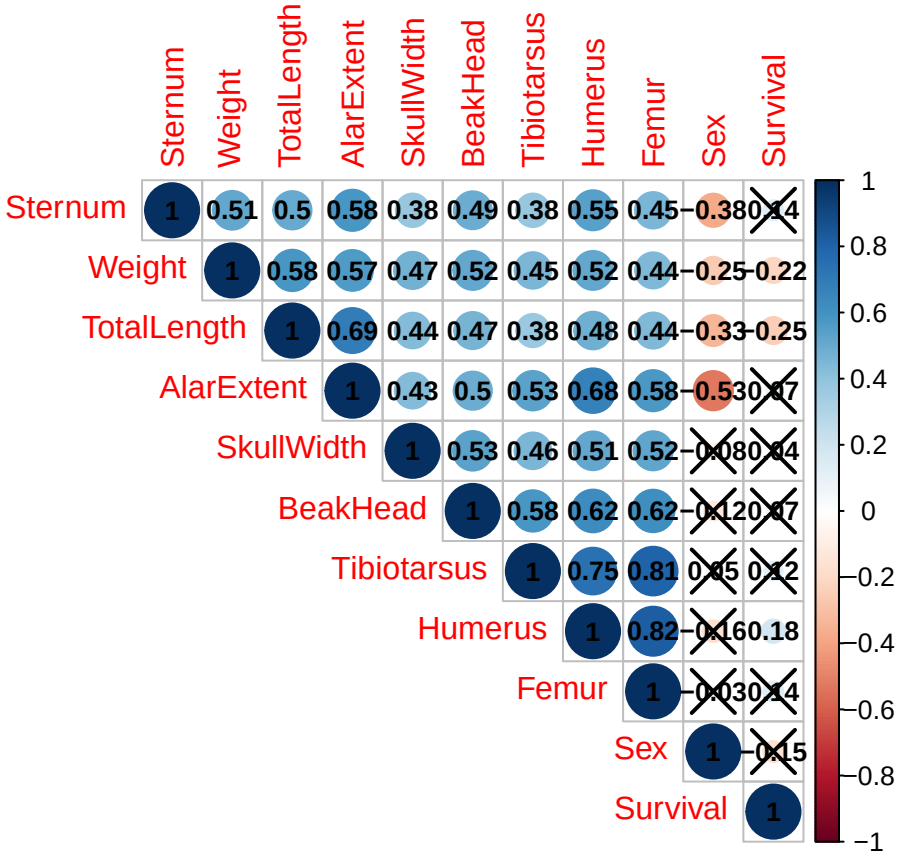
| Sex | Survival | TotalLenght (mm) | AlarExtent (mm) | Weight (g) | BeakHead (mm) | Humerus (inches) | Femur (inches) | Tibiotarsus (inches) | SkullWidth (inches) | Sterum (inches) |
|---|---|---|---|---|---|---|---|---|---|---|
| 0:87 | 0:64 | Min. :152.0 | Min. :230.0 | Min. :22.60 | Min. :29.80 | Min. :0.6590 | Min. :0.6530 | Min. :1.011 | Min. :0.5510 | Min. :0.7340 |
| 1:49 | 1:72 | 1st Qu.:157.0 | 1st Qu.:242.0 | 1st Qu.:24.57 | 1st Qu.:31.10 | 1st Qu.:0.7177 | 1st Qu.:0.7017 | 1st Qu.:1.112 | 1st Qu.:0.5920 | 1st Qu.:0.8090 |
| NA | NA | Median :160.0 | Median :246.0 | Median :25.55 | Median :31.60 | Median :0.7330 | Median :0.7130 | Median :1.133 | Median :0.6020 | Median :0.8410 |
| NA | NA | Mean :159.5 | Mean :245.2 | Mean :25.52 | Mean :31.57 | Mean :0.7319 | Mean :0.7130 | Mean :1.134 | Mean :0.6025 | Mean :0.8399 |
| NA | NA | 3rd Qu.:162.0 | 3rd Qu.:249.0 | 3rd Qu.:26.50 | 3rd Qu.:32.02 | 3rd Qu.:0.7482 | 3rd Qu.:0.7312 | 3rd Qu.:1.162 | 3rd Qu.:0.6110 | 3rd Qu.:0.8652 |
| NA | NA | Max. :167.0 | Max. :256.0 | Max. :31.00 | Max. :33.40 | Max. :0.7800 | Max. :0.7670 | Max. :1.230 | Max. :0.6400 | Max. :0.9270 |

Sex Column : 0 = Male, 1 = Female
Survival Column : 0 = Survived, 1 = Perished

## Correlation Map

A correlation map between all the variables was built as below. The positive and negative correlation are indicated by blue and red colors, with the correlation coefficient listed. 95 percent of significance level is used for the P value, the × indicates the correlation is not significant.

From the graph below, we can see the **Humerus** and **Tibiotarsus** are highly correlated with **Femur**, the **Weight** and **TotalLength** are negatively correlated with **Survival**, and **Humerus** is positively correlated with **Survival**



## Binary logistic regression

Binary logistic regression model method has been chosen to fitted the model to **Survival**. A link function has been used in the model on the response variable **Survival**. The logit link function is used to model the probability of 'success' as a function of covariates. The purpose of the logit link is to take a linear combination of the value and convert to the scale of a probability (between 0 and 1).

$$\mathbf{Logit}(\pi_{\mathbf{i}}) = \mathbf{Log}(\frac{\pi_{\mathbf{i}}}{\mathbf{1}-\pi_{\mathbf{i}}})$$

Two binary logistic models were fitted to *Survival*, one with interactions and one without any interactions. The model was initially reduced using *stepAIC* from *MASS* package. Then non significant variables have been dropped one by one, in the order from highest P-value to the lowest P-value. Both model equation includes only significant terms (P<0.05, Ward's test), the final two models have been given below:

**Model 1 : Without interaction terms**

$$\log\left[\frac{P(\widehat{\text{Survival}}=1)}{1-P(\widehat{\text{Survival}}=1)}\right] = 24.385 - 1.455(\text{Sex}_1) - 0.381(\text{TotalLength}) - 0.752(\text{Weight}) + 56.499(\text{Humerus}) + 17.835(\text{Sternum}) \tag{1}$$

**Model 2 : With Interaction Terms**

$$\begin{aligned}\log\left[\frac{P(\widehat{\text{Survival}}=1)}{1-P(\widehat{\text{Survival}}=1)}\right] = {} & -607.772 + 27.891(\text{Sex}_1) - \\ & 7.38(\text{TotalLength}) + 7.32(\text{AlarExtent}) + \\ & 1374.552(\text{Humerus}) - 1475.341(\text{Femur}) - \\ & 64.971(\text{Tibiotarsus}) - 275.871(\text{SkullWidth}) + \\ & 1098.86(\text{Sternum}) - 35.643(\text{Sex}_1 \times \text{Sternum}) + \\ & 20.598(\text{TotalLength} \times \text{SkullWidth}) - 6.707(\text{TotalLength} \times \text{Sternum}) - \\ & 12.209(\text{AlarExtent} \times \text{SkullWidth}) - 1186.967(\text{Humerus} \times \text{Tibiotarsus}) + \\ & 1312.952(\text{Femur} \times \text{Tibiotarsus})\end{aligned} \tag{2}$$

## Model Assumptions & Diagnostics

The function **anova()** was used to compare two models, the result shows that the model with extra interaction terms, with loosing 9 degree of freedom is not statistically better than the model without the interactions. The more simple model 1 is prefered for better interpretation and prediction.

Table 1: Analysis of Deviance Table

| Resid. Df | Resid. Dev | Df | Deviance | Pr(>Chi) |
|---|---|---|---|---|
| 130 | 133.6 | NA | NA | NA |
| 121 | 127.7 | 9 | 5.945 | 0.7454 |

Similarly, the drop in deviance test has also been conducted to compare both models, with similar result:

$$\mathbf{Pr}(\chi^{\mathbf{2}}_{\mathbf{9}} > (\mathbf{133.64} - \mathbf{127.70})) = \mathbf{0.745908}$$

In order to ensure the model 1 is adequate, the goodness of fit test has then been carried out on the model without interactions. The result of large $p$-value indicate that the model is appropriate and adequate.

$$\mathbf{Pr}(\chi^2_{\mathbf{130}} > \mathbf{133.64}) = \mathbf{0.3955299}$$

Therefore we can conclude the Model 1 will be our final model.

## Prediction / Visualisaiton

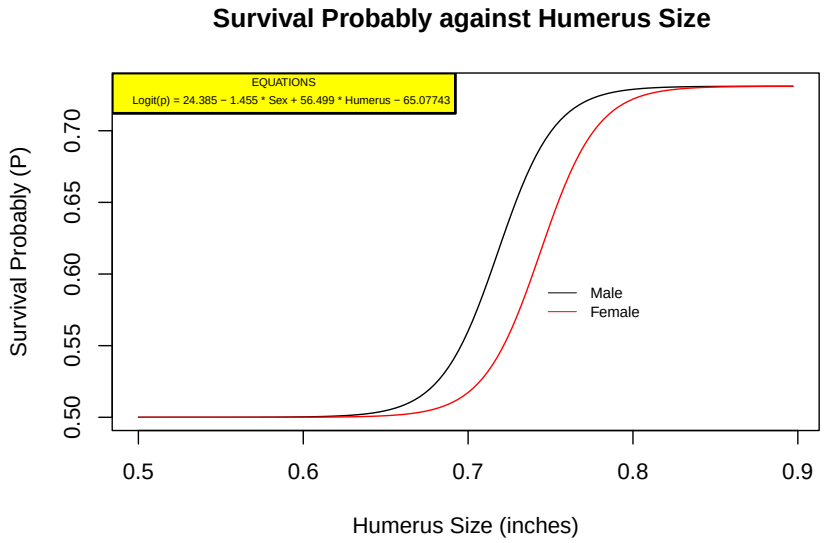
In order to compare the model predicted value with the survival records in our original data. The function **predict()** has been used, after the reverse transformation of the logit function, a list of **survival probability ( between 0 and 1)** was produced.

To conduct a meaningful comparison between original survival data records **(0:Perished, 1:Survived)**, with our list of **survival probability**. A classification was added, to define that the **probability >= 0.5** will be considered **survived**,the **probability < 0.5** will be considered **perished** . The **probability = 0.5** can be classified in both categories, however due to the decimal precision, our result does not include exact 0.5 situation.

Below table is the result of the pair-wise comparison between predicted value and original data. There are **46 Perished** cases predicted correctly, and **58 Survived** cases predicted correctly. The model has correctly predicted **76.47%** by feeding the original data, a total of 136 records.

| | Perished | Survived |
|---|---|---|
| Perished | 46 | 14 |
| Survived | 18 | 58 |

To visualize the predicted survival probability trend with just the increase of Humerus size, we've also taken the mean values of Weight, Sternum, and Total Length from the original data as three constant values for the model prediction, then 136 evenly distributed the data points were generated for Humerus Size, ranging from 0.5 inches to 0.9 inches, with steps of 0.0029411. Two lines were plotted respectively for male(0) and female(1) below. From the graph it shows the probably of survival has increased along with the increase of Humerus size.



## 4 Discussion

From the final model equation, note that the main effects of Total Length, Weight are negative, but the Humerus and Sternum are positive. The model interpretation is as follows:

1. The odds of a male house sparrow surviving were about 5 times (exp(-1.455)) the odds of a female house sparrow surviving.
2. With 1 unit (mm) decrease in Total Length, the odds of house sparrow surviving will be approximately increased by 46 percent.

   With 1 unit (gram) decrease in Weight, the odds of house sparrow surviving will be approximately doubled (212 percent).
3. With 0.01 unit (inches) increase in Humerus, the odds of house sparrow surviving will be approximately increased by 76 percent.

   With 0.01 unit (inches) increase in Sternum, the odds of house sparrow surviving will be approximately increased by 20 percent.

Our results are slightly different to the result produced by others. However, the species, location, season may also contribute influences on the result.

We note that there are nearly twice the male house sparrows than female sparrows in our data, the model may be improved with more evenly distributed gender cases.

Apart from the body measures, the **age** could also be a strong factor affecting the surviving ability in severe weather, also the wing and tail asymmetry may contribute to the balance control in strong winds. The two heaviest birds (outliers) weighing over 30g were both perished in our records.

# References

Brown, Charles R., Brown, Mary Bomberger. 1998. "Intense Natural Selection on Body Size and Wing and Tail Asymmetry in Cliff Swallows During Severe Weather." *Evolution* 52(5): 1461–75. https://doi.org/10.1007/BF01237725.

Pugesek, Bruce H., Tomer, Adrian. 1996. "The Bumpus House Sparrow Data: A Reanalysis Using Structural Equation Models." *Evolutionary Ecology*, 1387–1404. https://doi.org/10.1007/BF01237725.