Bumpus' House Sparrows Survival: A Binomial Regression Approach

Saif Ali Athyaab - 23810756

1 Executive Summary

This report is on modelling the survivability of Bumpus' house sparrows based on some physical measurements. A total of 136 data records were available, including the identifier of the bird, sex of the bird, survivability of the bird, length from beak to tail, length of extended wings, weight of the bird, length of the beak and head, length of humerus, length of femur, length of tibiotarsus, skull width, length of keel of sternum. The model for a reduced binomial logistic regression was the best. A model with no interaction terms was preferred over one with 2 interactions between sex and continuous variables. Although the latter was marginally better from a statistical point of view, the results from a likelihood estimation test concluded otherwise. The selected model showed that the survival had a negative correlation between total length and weight of the bird, and was positive for Beak head length. Humerus length and sternum showed exponential positive correlation with survival. On the whole, Males showed better odds of survival.

2 Introduction

The dataset consists of observations recorded by Hermon C. Bumpus, a zoologist and professor of comparative anatomy at Brown University, in 1898[1]. On the night of February 1 1898, he received a collection of 136 birds (of which 72 survived) after a storm had hit Rhode Island. 9 morphological measures were collected along with the Sex of the bird. The sheer severity coupled with the prolonged nature of the storm was a rare event which made him think about a differential survival [1] of the birds received. Theories like 'evolution' and 'survival of the fittest' prompted subsequent studies on the same dataset using path analysis [2], structural equation models [4], geographic variations [6,7], double exponential [9] and logistic regression [11]. Some of these methods involve the creation of new variables such as fitness to model interaction effects.

As linear regression methods were the norm, it was a breakthrough when Janzen and Stern [11] introduced the logistic model to effectively represent the dichotomous nature of Survival (TRUE or FALSE). They regressed survivorship on logarithms of total length, wing extent, skull width, head length, humerus length, femur length, tibiotarsus length, sternum length, and one-third the logarithm of weight [11].

In this report we estimate the survival using a binomial model. This report is as follows. In the next section we describe the statistical methodology used, followed by the results section. This is followed by a discussion of the findings, which are compared with the literature discussed in the introduction.

3 Methodology

We will commence the exploration of data by descriptive statistics and examination of any outliers. Following this a binomial statistical model will be fitted to the data with the Survival as response. Interaction terms will also be included, and any appropriate transformation of data will be investigated. The model will be reduced to significant terms only. A simple model that is easier to interpret is preferred to a more complex model that may be better from a statistical point of view. The final model will be interpreted to explain the difference of Survival on the variables in the data. Note that the continuous variables in the data had different units for length such as mm and inches. However, no single-scale unification has been carried out in view of simpler and familiarity of inference. All statistical analysis will be conducted in the R Studio statistical environment. Statistical significance will be taken at $\alpha = 0.05$ (5%). The metrics used to measure the suitability of the model include Residual deviance, Likelihood Ratio Test, and the AIC comparison.

4 Results

The analysis can be divided into 2 main parts. Part 1 discusses initial exploratory data analysis followed by Part 2 which is the actual model fitting.

4.1 Part 1

During the exploratory phase, the main focus was to get a thorough understanding of the data itself. The ID variable which represented the unique identifier was discarded. Also, no missing or invalid values were found.

Table 1 – Dataset Descriptive Statistics

Variable	***Description***	***Summary Statistics***
Sex	A 2-level factor variable representing Male or Female	Male(m) = 87, Female(f) = 49
Survival	A binary variable informing survivorship of the bird	TRUE = 72, FALSE = 64
Total Length (mm)	Tip of the beak to the tip of the tail	152 – 167, mean = 159.5
Alar Extent (mm)	Tip to tip of the extended wings	230 – 256, mean = 245.2
Weight (g)	Weight of the bird	22.6 – 31.0, mean = 25.52
BeakHead (mm)	Length of beak and head, measured from tip of the beak to the occiput	29.8 – 33.4, mean = 31.57
Humerus (inches)	Length of humerus	0.659 – 0.78, mean = 0.7319
Femur (inches)	Length of femur	0.653 – 0.7670, mean = 0.713
Tibiotarsus (inches)	Length of Tibiotarsus	1.011 – 1.23, mean = 1.134
SkullWidth (inches)	Width of skull measured from the postorbital bone of one side to the postorbital bone of the other	0.551 – 0.64, mean = 0.6025
Sternum (inches)	Length of keel of sternum	0.734 – 0.927, mean = 0.8399

Weight was found to be moderately right-skewed and Tibiotarsus was moderately left-skewed through histograms. Two Weight outliers from boxplot, namely the 46th and 110th, weighed 1.5g higher than average.

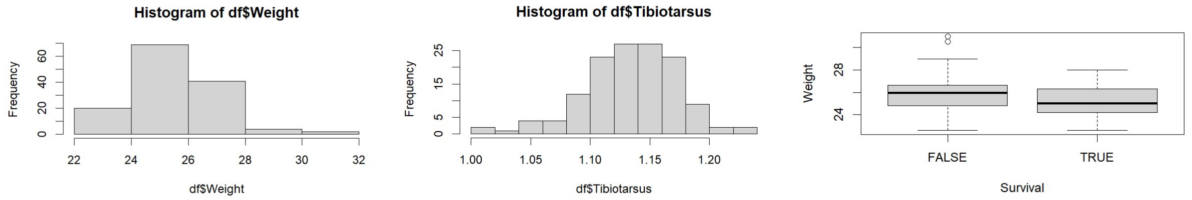

Figure 1 – Histograms of Weight and Tibiotarsus variables, Boxplot of Weight against Survival

4.2 Part 2

Given the binomial nature of the selected response variable, it was deemed appropriate to fit a Generalized Logistic Model of the Binomial family with the Logit link function to map the probabilities of survivorship to the covariates.

4.2.1 Full Standard Binomial Regression

Model fitting commenced by modelling all variables against the Survival. Reference level for Sex is Male(m). The fitted model equation in this case is:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = 10.83 - 1.64\widehat{Sexm} - 0.43\widehat{TotalLength} + 0.02\widehat{AlarExtent} - 0.86\widehat{Weight} + 0.48\widehat{BeakHead} + 40.85\widehat{Humerus} - 2.72\widehat{Femur} + 6.89\widehat{Tibiotarsus} + 13.74\widehat{SkullWidth} + 16.25\widehat{Sternum}$$

4.2.2 Reduced Standard Binomial Regression

Next, StepAIC was used to iteratively reduce the full model into one consisting of mostly statistically significant terms. This was selected as our final model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = 15.85 + 1.59\widehat{Sexm} - 0.42\widehat{TotalLength} - 0.83\widehat{Weight} + 0.61\widehat{BeakHead} + 51.4\widehat{Humerus} + 16.82\widehat{Sternum}$$

4.2.3 Interaction Terms

A two-way interaction between all variables was modelled resulting in 22 combinations predictors. StepAIC was again employed to intelligently reduce the predictors to just 8. However, the residual deviance remained in the same range and these interaction models were discarded. Sex with Sternum and TotalLength were the only interactions retained. A likelihood ratio test with the prior model yielded 0.06976 p-value indicating there was negligible significant interaction effect from the Sex variable.

4.2.4 Weight and Tibiotarsus Transformations

Drawing from the moderately skewed behaviour from the Part 1 EDA, Weight was transformed to a square root and Tibiotarsus was transformed to a cube, but the model was discarded due to having few significant terms. StepAIC on this model also removed the cube transformation.

4.2.5 Weight Outliers Removal

All the above regressions were repeated after removal of observation 46 and 110 which had weights of 30.5 and 31. These were above the 3rd IQR. The residual deviance and AIC values were no more than 5% different from the full dataset regressions and hence this idea was dropped to ensure data integrity.

4.2.6 Models Metrics Comparison

2 main metrics are used to compare the different models, the first being Residual Deviance (lower is better) and AIC (lower is better). The Null Model deviance (only intercept) was 188.07 for full dataset and 185.02 for outlier treated dataset.

Table 2 - Model Comparison

Model Type	**Residual Deviance**	**Goodness of Fit (AIC)**	**Number of Covariates excluding Intercept**
Full Model	130.21	152.21	10
Reduced Model	131.57	145.57	6
Reduced Model with Sex interaction terms	128.3	144.3	7
Reduced Transformed variables Model	131.56	145.56	6
Reduced Outlier treated Model with Sex:TotalLength interaction term	130.33	144.33	6

5 Discussion

From the Reduced model equation (4.2.2), all the variables in the data except BeakHead are significant predictors of the Survival. The effects of the variables on the exponentials of covariates are as follows.

1. For the binary variable Sex we can say the odds of survival for Males are estimated to be exp(1.5958) = 4.93 times the odds of survival for Females. Males are 5x more likely to survive with the Confidence Interval (CI) being 1.75 to 13.88 times.
2. Total Length is negatively correlated to Survival with a 1 mm increase in length reducing the Survival with a multiplicative rate of 0.65 lying between a CI of 0.54 and 0.80
3. Weight is negatively correlated to Survival with a 1g increase in weight reducing the Survival with a multiplicative rate of 0.44 lying between a CI of 0.27 and 0.70
4. Beak Head is positively correlated with Survival with an odds of 1.84 and a CI of 0.8 and 4.25.
5. Length of Humerus and Length of Sternum have exponentially high positive correlation to Survival. Humerus leads with 2e+22 odds between a CI of 5e+09 and 8.8e+34. Sternum is slightly behind with an odds of 2e+07 between a CI of 2.2 and 1.85e+14

Our model is similar to that of Janzen et. al [11]. However, we have also well-explored all other covariate possibilities including interaction terms, transformations to address skewness, and outlier treatment before arriving at a reduced binomial logistic model. We have chosen a simple model, rather than a complex one involving functional transformations, for ease of inference.

6 References

1. Johnston, R. F., Niles, D. M., & Rohwer, S. A. (1972). Hermon Bumpus and Natural Selection in the House Sparrow Passer domesticus. Evolution, 26(1), 20–31. https://doi.org/10.2307/2406980
2. Crespi, B. J., & Bookstein, F. L. (1989). A Path-Analytic Model for the Measurement of Selection on Morphology. Evolution, 43(1), 18–28. https://doi.org/10.2307/2409161

3. Buttemer, W. A. (1992). Differential Overnight Survival by Bumpus' House Sparrows: An Alternate Interpretation. The Condor, 94(4), 944–954. https://doi.org/10.2307/1369291

4. Pugesek, B.H., Tomer, A. (1996). The Bumpus house sparrow data: A reanalysis using structural equation models. Evol Ecol 10, 387–404. https://doi.org/10.1007/BF01237725

5. Crespi, B. J. (1990). Measuring the effect of natural selection on phenotypic interaction systems. The American Naturalist, 135(1), 32-47. https://www.journals.uchicago.edu/doi/epdf/10.1086/285030

6. Johnston, R. F., & Selander, R. K. (1973). Evolution in the house sparrow. III. Variation in size and sexual dimorphism in Europe and North and South America. The American Naturalist, 107(955), 373-390. https://www.journals.uchicago.edu/doi/epdf/10.1086/282841

7. Johnston, R. F., & Selander, R. K. (1971). Evolution in the House Sparrow. II. Adaptive Differentiation in North American Populations. Evolution, 25(1), 1–28. https://doi.org/10.2307/2406496

8. Schluter, D. (1988). Estimating the form of natural selection on a quantitative trait. Evolution, 42(5), 849-861. https://doi.org/10.1111/j.1558-5646.1988.tb02507.x

9. Manly, B. F. J. (1976). Some examples of double exponential fitness functions. Heredity, 36(2), 229-234. https://doi.org/10.1038/hdy.1976.27

10. Lowther, P. E. (1977). Selection Intensity in North American House Sparrows (Passer domesticus). Evolution, 31(3), 649–656. https://doi.org/10.2307/2407530

11. Janzen, F. J., & Stern, H. S. (1998). Logistic regression for empirical studies of multivariate selection. Evolution, 52(6), 1564-1571. https://doi.org/10.1111/j.1558-5646.1998.tb02237.x