STAT2402 ASSIGNMENT 2

Student Name: Aaminah Irfan Student Number: 23642166

Executive Summary:

The primary aim of this report is to develop a simplified logistic linear model to understand the relationship between the survival rates of English sparrows and their morphological attributes using Bumpus 1898 data set. The dataset includes information on 136 sparrows, featuring continuous morphological characteristics (Total Length, Alar Extent, Weight, Beak to Head Length, Length of Humerus , Femur, Tibiotarsus, Skull Width and Sternum), gender (Sex), and a binary survival variable (Survival). Logistic models, specifically Generalized Linear Models (GLMs), are employed to investigate how these variables influence survival. After thorough data exploration, variable selection, and model building, we it was found that only Total Length, Weight, and Humerus exhibit statistical significance in relation to survival rates. These three variables are retained for further analysis. Subsequently, we construct models to explore potential interactions among the explanatory variables and their impact on survival rates. Model 2 is chosen as the final model based on likelihood ratio tests (lrtest) and Pearson's Residual plots, indicating its superior fit and simplicity. Model 2 reveals that as Total Length and Weight increases, survival rates for both Male and Female decreases. Alternatively as Humerus Length increases survival rate for both gender increases. Finally it was revealed that Females had lower odds of survival compared to Males.

Introduction:

On February 1, 1898, a severe snow- storm hit New England [1]. A number of English sparrows, presumably exhausted due to the storm, were brought to the Anatomical Laboratory of Brown University for analysis [1]. Out of many birds, only a few surviving ones remained, and the rest died. Bumpus (1898) measured nine physical characteristics along with Gender [1]. Multiple studies have been conducted using the data set by Bumpus, to examine the relationship between survival and natural selection and / or morphological characteristics. One notable re-analysis (Peter O'Donald, 1973) [1] aimed to delve deeper into the intensity of natural selection and its relation to survival rates of English sparrows by using advanced statistical techniques, including structural equation modelling (SEM). Results from the study revealed that survival rate significantly increased with increasing general size of the sparrows and leg size and head size was unrelated to the survival rate. Wing Length, independent to its relationship with general size, was also significantly correlated with survival rate where sparrows with shorter wings have higher survival rates. It was also found that Males generally had higher survival rates than females. Another study (William A. Buttemer) [3] provided an alternate interpretation of survival pattern reported by Bumpus by conducting a study of morphometric and energy relation on survival patterns for sparrows. Results revealed that birds were unable to survive due to cold stress, severe winter storms and had little relation with morphological characteristics. This research paper aims to devise a more simple and practical method using logistic linear models for estimating the relationship between survival rates and morphological characteristics of English sparrows. Data set used had information on 136 sparrows and its morphological characteristics. All variables in the data are continuous and numerical variables, except 'Sex' which is a categorical variable and 'Survival' which is a logical variable. Table 1 provides the summary of data.

Table 1: Bumpus Data Summary

Variable	**Number of Observations**	**Value Range**	**Mean**
ID: an identifier of bird	136	1-136	68.50

Sex: sex of the bird (m = Male, f = Female)	m = 87 f = 49		
Survival: a logical variable indicating survival status of the bird (**T** = Survived, **F** = Died)	False = 64 True = 72		
TotalLength: measured from tip of the beak to the tip of the tail (mm)	136	152.00 - 167.00	159.50
AlarExtent: measured from tip to tip of the extended wings (mm)	136	230.00 - 256.00	245.20
Weight: weight of the bird (g)	136	22.60 - 31.00	25.52
BeakHead: length of beak and head, measured from tip of the beak to the occiput (mm)	136	29.80 - 33.40	31.57
Humerus: length of humerus (inches)	136	0.66 - 0.78	0.73
Femur: length of femur (inches)	136	0.65 - 0.77	0.71
Tibiotarsus: length of tibiotarsus (inches)	136	1.01 - 1.23	1.13
SkullWidth: width of skull measured from the postorbital bone of one side to the postorbital bone of the other (inches)	136	0.55 - 0.64	0.60
Sternum: length of keel of sternum (inches)	136	0.73 - 0.93	0.84

Methodology:

Data was explored using graphical and numerical summaries and variables were converted to appropriate number type. Irrelevant variables were excluded from the data and final ones were used for further statistical analysis. To examine the relationship between the selected variables and survival rates, logistic models (or Generalised Linear Models (GLMs) in R) were employed where the outcome represents survival (1 for survival, 0 for non-survival). Response variable was Survival and explanatory variables were Sex, Total Length, AlarExtent, Weight, Beakhead, Humerus, Femur, Tibiotarsus, SkullWidth and Sternum. For all statistical test, significance level of 5% was used. To assess the significance of the variable(s) or the model(s), following Null and Alternate hypothesis was used:

Null hypothesis: There is no relationship between the explanatory variable(s) and Survival
Alternate Hypothesis: There is relationship between the explanatory variable(s) and Survival.

If the p-value is less than 5%, null hypothesis would be rejected. Alternatively, if p-value is greater than 5% we fail to reject the null hypothesis. We model the probability of survival, denoted as π_i each individual sparrow i as follows:

$$[\pi_i = \frac{e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+...+\beta_p X_{pi}}}{1+e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+...+\beta_p X_{pi}}}]$$

Maximum Likelihood Estimators (MLEs) for binary responses choose the parameters of the model such that the data has the highest probability of occurrence. For large enough sample size, MLEs are: (i) Essentially unbiased, (ii) Have obtainable standard errors, (ii) Are efficient: more precise than other estimators, (iv) Have an approximate normal distribution.

Firstly, best model for each continuous variable was chosen. To select individual models for each continuous variable, we started by performing regression analyses. In these analyses, we used Survival as the response variable and included the continuous variable along with Sex as additional variables. This resulted in a full model that encompassed Survival, the continuous variable, Sex, and an interaction term between Sex and the continuous variable. Likelihood ratio test (lrtest) was then employed to compare this full model with a model that excluded the interaction term. Subsequently, final model was compared with a model that did not include the Sex variable. The final model was selected for each continuous variable based on these comparisons. Secondly, continuous variables that did not show statistical significance in regression models, whether with or without interactions and additional explanatory variables, were excluded from the subsequent analysis steps. R regression analyses on the remaining continuous variables and the categorical variable (Sex) was employed to investigate their relationships with Survival. This included examining all possible interactions between these variables. Stepwise Akaike Information Criterion

(stepAIC) was utilised to select the most suitable model. Subsequently, Model Selection techniques were applied to refine the model by eliminating statistically insignificant variables. The final model was determined by comparing it with alternative models using the likelihood ratio test (lrtest), and the best-fitting model was chosen.

Results

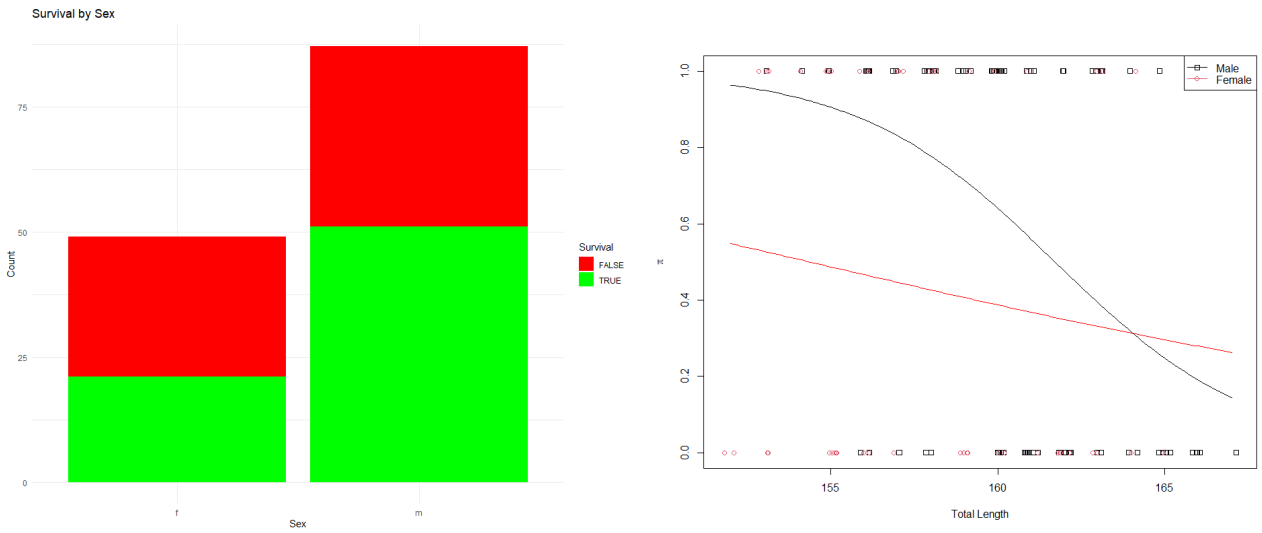
During data analysis, Survival variable was converted to 'numeric' type from 'logical' so the probability of Survival can be modeled on the scale of 0 to 1 . Similarly 'Sex' variable was converted to 'numeric' type where 'Male' was denoted 0 and 'Female' was denoted '1'. Furthermore, 'ID' variable was removed from the data set as it is an unique identifier of each observation and including it would not contribute the understanding the main aims of the analysis. Best fitting models for each continuous variable was tested, including interaction terms and relationship with Sex, and chosen for further analysis. Best-fitting Individual Models for each continuous variable are shown in Table 2.

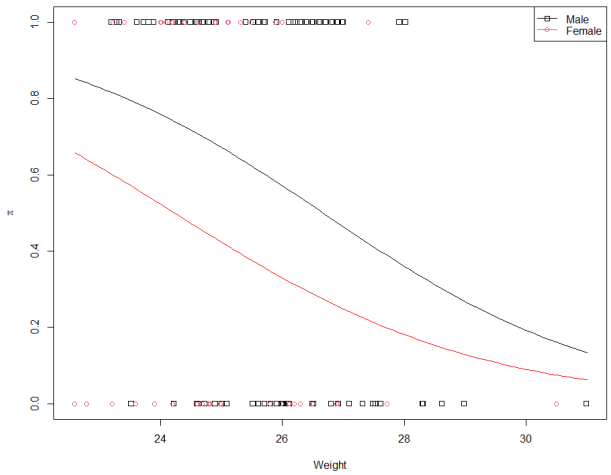
Table 2: Continuous Variables' Best Fitting Model

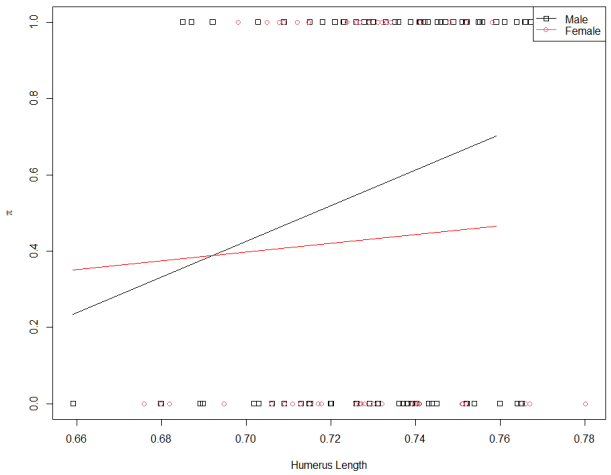
Variable	Best-Fitting Model
TotalLength	$\text{logit}(\pi_i) = 54.49 - (0.34 \cdot \text{TotalLength}) - (41.92 \cdot \text{SexN}) + (0.26 \cdot \text{TotalLength:SexN})$
AlarExtent	$\text{logit}(\pi_i) = -5.93 + (0.02 \cdot \text{AlarExtent})$
Weight	$\text{logit}(\pi_i) = 11.19 - (0.42 \cdot \text{Weight}) - (1.02 \cdot \text{SexN})$
Beakhead	$\text{logit}(\pi_i) = -5.90 + (0.19 \cdot \text{BeakHead})$
Humerus	$\text{logit}(\pi_i) = -12.03 + (16.60 \cdot \text{Humerus})$
Femur	$\text{logit}(\pi_i) = -8.09 + (11.51 \cdot \text{Femur})$
Tibiotarsus	$\text{logit}(\pi_i) = -6.53 + (5.87 \cdot \text{Tibiotarsus})$
SkullWidth	$\text{logit}(\pi_i) = -2.72 + (4.71 \cdot \text{SkullWidth})$
Sternum	$\text{logit}(\pi_i) = -5.79 + (7.05 \cdot \text{Sternum})$

Among the models examined, only TotalLength, Weight, and Humerus displayed statistical significance, and these three variables were selected for further analysis. Plot for significant variables are shown in Table 3. The remaining continuous variables did not exhibit statistical significance in any model configuration, including those with and without interaction terms and additional Sex variable.

Table 3: Plots of Significant Variables







To explore potential interactions among remaining significant explanatory variables and their relationship with survival rates, we conducted regressions involving the selected continuous variables and the Sex variable, considering all possible interactions between them. Subsequently, we employed Stepwise Akaike Information Criterion (stepAIC) to identify the optimal model. The resulting Model 1 can be represented as:
Model 1:

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = 44.06 - (0.46 \cdot \text{TotalLength}) - (0.64 \cdot \text{Weight}) + (64.20 \cdot \text{Humerus}) - (41.91 \cdot \text{SexN}) + (0.25 \cdot \text{TotalLength} : \text{SexN})$$

However, Model 1 still included some insignificant variables. To address this, we performed Backward Model Selection, resulting in the final Model 2:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = 23.04 - (0.33 \cdot \text{TotalLength}) - (0.63 \cdot \text{Weight}) + (63.83 \cdot \text{Humerus}) - (1.68 \cdot \text{SexN})$$

Comparing Model 1 (AIC: 147.45) and Model 2 (AIC: 149.02) using the likelihood ratio test (lrtest) indicated a p-value greater than 0.05, signifying that Model 2 outperformed Model 1. To make a final selection between the two models, we examined Pearson's Residual plots (Figure 1,2).

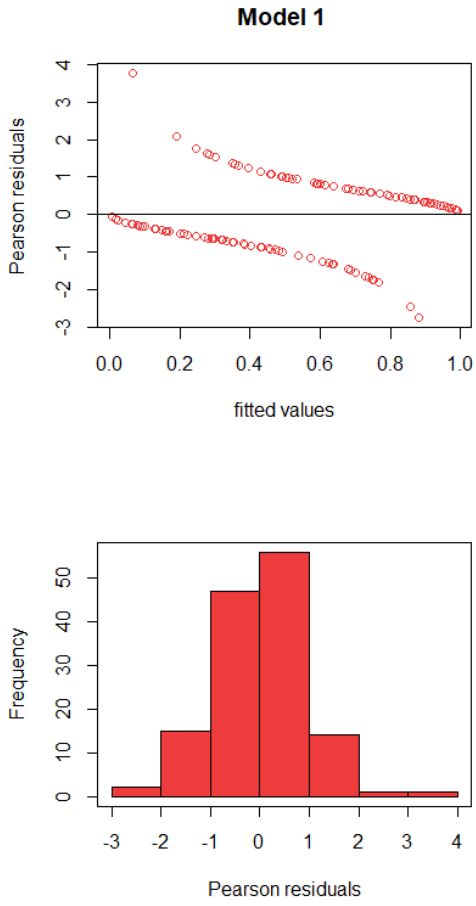


Figure 1: Pearson's Residuals Plot

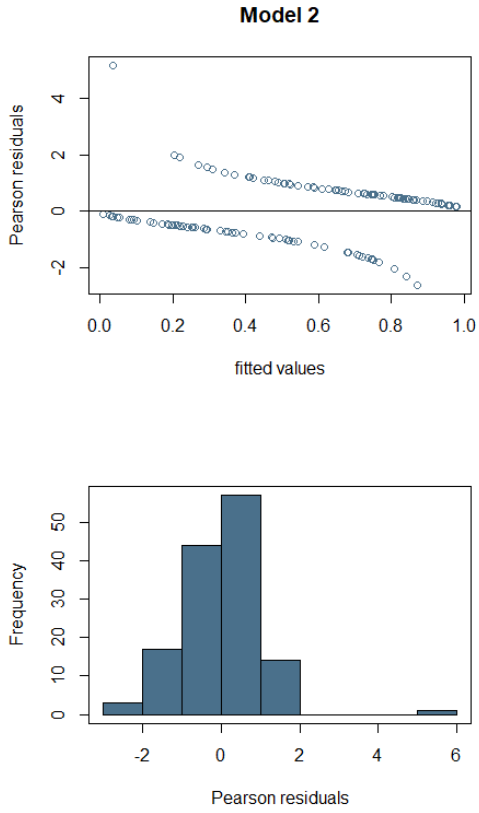


Figure 2: Pearson's Residuals Plot

The Pearson's Residual Plot for Model 2 appears superior to that of Model 1 since it exhibits a distribution closer to a mean of 0, suggesting that residuals are centered around this value. However, when we examine the histogram of Pearson's Residuals for Model 1, we observe a distribution that more closely resembles normality. Conversely, in the case of Model 2, there appear to be outliers. Several factors contribute to the selection of Model 2 as the final choice: (i) it offers greater simplicity, (ii) it provides a better overall fit to the data, and (iii) the Pearson's Residuals for Model 2 exhibit a closer alignment with a mean of 0, (iv) all coefficients are statistically significant.

Discussion

Final Model chosen was Model 2 compared to Model 1 due to its statistical significance and simplicity. The effects of the explanatory variables on log-odds of survival for sparrows are as follows:

1. For continuous variable Total Length, for one unit rise in TotalLength, with all other explanatory variables fixed, will result in $exp(\beta_1 = -0.33) = 0.72$ multiplicative change in the odds of survival of sparrows. As males and females' Total Length increases, the odds of survival reduce.
2. For continuous variable Weight for one unit rise in Weight, with all other explanatory variables fixed, will result in $exp(\beta_2 = -0.63) = 0.53$ multiplicative change in the odds of survival of sparrows. As males and females' Weight increases, the odds of survival reduce.
3. For continuous variable Humerus for one unit rise in Humerus, with all other explanatory variables fixed, will result in $exp(\beta_3 = 63.83) = 5.27 \times 10^{27}$ multiplicative change in the odds of survival of sparrows. As males and females' length of Humerus increases, the odds of survival increase.

4. For binary variable Sex, with all other explanatory variables fixed, we can say the odds of survival for females are estimated to be $exp(\beta_4 = -1.68) = 0.185$ times the odds of survival for males. Females are less likely to survive than Males.

The relationships can be seen in Figure 1 where as Total Length increases, survival rate decreases for both Males and Females. Similarly, as Weight increases survival rate for both Males and Females decreases at almost the same rate. Alternatively, as Humerus Length increases, survival rate for Male and Female increases, where Male survival rate increases at a faster rate than for females. For all significant continuous variables, female survival rate stays consistently lower than males.

Despite the strength and simplicity of out model, it is important to note potential weaknesses and alternative approaches. Firstly, sample size of male and female differed substantially where the data set contained information on 38 more males than females. This can be seen in Figure 1 and Table 1 , where there is more data on Male sparrows than Female sparrows. This may have resulted in bias in the analysis of survival rates. Secondly, while the analysis only focused on relationship between morphological measurements and the survival rate, there ma be other unmeasured factors that could influence survival rates in English sparrows such as extreme weather conditions as explored by Buttemer,1992 [3].

References

[1] O'Donald P. (1973). A FURTHER ANALYSIS OF BUMPUS' DATA: THE INTENSITY OF NATURAL SELECTION. *Evolution; international journal of organic evolution*, *27*(3), 398–404. https://doi.org/10.1111/j.1558-5646.1973.tb00686.x (O'Donald 1973)

[2] Pugesek, B. H., & Tomer, A. (1996). The Bumpus house sparrow data: A reanalysis using structural equation models. *Evolutionary Ecology*, *10*(4), 387–404. https://doi.org/10.1007/bf01237725

[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

O'Donald, Peter. 1973. "A FURTHER ANALYSIS OF BUMPUS' DATA: THE INTENSITY OF NATURAL SELECTION." *Evolution* 27 (3): 398–404. https://doi.org/10.1111/j.1558-5646.1973.tb00686.x.