STAT2402 Analysis of Observations Assignment 2

Himakar Gadham - 23783777

Executive Summary

In our binomial regression analysis examining the survival of English sparrows, several morphological characteristics stood out as statistically significant predictors. For instance, TotalLength was negatively associated with survival rates, implying that longer birds might be more vulnerable, possibly due to elevated energy requirements or increased visibility to predators. On the other hand, AlarExtent, representing wing length, was positively correlated with survival, suggesting benefits in terms of flight efficiency and evasion from predators. Weight also emerged as a significant factor, with lighter birds showing higher odds of survival, which may be attributed to greater mobility and energy efficiency. While the Humerus and Sternum measurements were significant in our final model, the specific biological mechanisms behind their influence on survival remain to be elucidated. Interestingly, other morphological features like SkullWidth and Femur did not show a statistically significant impact on survival, suggesting these may not be crucial factors in survival outcomes. These results offer preliminary insights into the factors affecting English sparrow survival and underscore the need for further studies that integrate ecological considerations and additional variables like behavior and environment.

Introduction

This report offers a thorough examination of a dataset acquired from the Anatomy Laboratory at Brown University, Rhode Island, captured on February 1, 1898, in the wake of a severe storm. The dataset, colloquially known as the Bumpus dataset, provides comprehensive morphological measurements along with survival statuses for a sample of English sparrows. By employing an array of statistical models—including linear, binomial, Poisson, and Gaussian regression models—and exploring various data transformations, we seek to understand how these morphological attributes correlate with the chances of survival for these birds. The insights derived from this study could serve as critical contributions to broader ecological research.

Dataset Description

The dataset encompasses multiple variables, each catering to different aspects of avian morphology and survival. These variables are as follows:

- **ID:** A unique identifier for each specimen.
- **Sex:** Gender of the bird, represented as 'm' for Male and 'f' for Female.
- **Survival:** A binary metric indicating whether the bird survived (TRUE) or perished (FALSE).
- **TotalLength:** The length from the beak's tip to the tail's tip, measured in millimeters.
- **AlarExtent:** The wingspan, calculated from one wingtip to the other, also in millimeters.
- **Weight:** The bird's weight, measured in grams.
- **BeakHead:** The combined length of the beak and head, from beak tip to occiput, in millimeters.
- **Humerus, Femur, Tibiotarsus:** Lengths of these specific bones, measured in inches.
- **SkullWidth:** The width of the skull, spanning from one postorbital bone to the other, in inches.
- **Sternum:** The length of the keel of the sternum, measured in inches.

Aim of the Analysis The principal objective of this analytical endeavor is to investigate how various morphological factors may influence the survival probabilities of English sparrows. Through data exploration, statistical modeling, and interpretive analysis, we aim to discern patterns and relationships among the dependent and independent variables. The forthcoming sections will outline the methodology employed, showcase the results obtained, and discuss their broader implications.

Methodology We will first explore the data by generating numerical summaries like mean, median, and standard deviation, as well as graphical summaries via histograms, scatter plots, and box plots. Following this, a generalized linear model (GLM) will be fitted to the data, incorporating a range of morphological and environmental predictors. Interaction terms between significant variables will be included to account for synergistic effects. The dataset will be examined for normality and linearity, and if required, appropriate transformations will be applied. A stepwise regression technique will be used to eliminate insignificant terms, leading to a more parsimonious model. While statistical robustness is important, a preference will be given to simpler models for the sake of interpretability. The final model's goodness-of-fit will be assessed using criteria like Residual Deviance and the Akaike Information Criterion (AIC). Ultimately, the final model will be interpreted to elucidate the relationship between the predictors and sparrow survivability. All statistical analysis will be conducted in the R statistical environment [9].

Results - The results section will include data exploration findings, tables, and graphs summarizing the dataset's characteristics. We will also describe the model fitting procedure and present the different model's coefficients.

Summary Statistics Here are the summary statistics for some of the key variables in the dataset:

Variable	Category	Frequency	Survival... False	Survival... True
Sex	f	49	28	21
	m	87	36	51
Survival		NA	64	72
Total		136	64	72

Variable	Mean	Median	Std..Dev.	Min	Max
TotalLength	159.5441176	160.000	3.5608315	152.000	167.000
AlarExtent	245.1911765	246.000	5.5210234	230.000	256.000
Weight	25.5250000	25.550	1.4752150	22.600	31.000
BeakHead	31.5742647	31.600	0.7023765	29.800	33.400
Humerus	0.7319412	0.733	0.0230782	0.659	0.780
Femur	0.7130294	0.713	0.0241133	0.653	0.767
Tibiotarsus	1.1335662	1.133	0.0407445	1.011	1.230
SkullWidth	0.6024779	0.602	0.0149958	0.551	0.640
Sternum	0.8399338	0.841	0.0396492	0.734	0.927

Survival Analysis The t-test analysis reveals that 'TotalLength' is a significant morphological parameter influencing both survival and sex in sparrows. Statistically significant differences were observed in 'TotalLength' between surviving and non-surviving sparrows, with non-survivors having a higher average length. Likewise, a significant difference in 'TotalLength' was observed between male and female sparrows, with males being larger on average. Both findings were supported by p-values well below the 0.05 alpha level and 95% confidence intervals that did not include zero, underlining the statistical validity of these conclusions.

The Chi-Squared test was used to assess the association between the survival outcome and the sex of sparrows. Despite obtaining a Chi-Squared value of 2.5257 with 1 degree of freedom, the p-value of 0.112 exceeded the standard alpha level of 0.05. Therefore, the test did not yield statistically significant results to reject the null hypothesis, indicating that there is insufficient evidence in the dataset to assert a significant relationship between the survival of sparrows and their sex.

A comprehensive correlation analysis across Pearson, Spearman, and Kendall coefficients revealed nuanced relationships among morphological variables. The most consistent correlation across all sets was between Humerus and Femur, ranging from moderate to high, highlighting their interconnected nature. Caution

against multicollinearity is advised against Pearson set especially as strong correlations were observed among Femur, Humerus, and Tibiotarsus. The Kendall set generally showed weaker correlations, suggesting less potent linear associations. Variability in correlations like that between TotalLength and AlarExtent across different sets implies that these relationships might be influenced by specific conditions or subsets.

Logistic Regression The application of multiple regression models, including logistic(Model 2 and 3) and linear regressions(Model 1) as well as Gaussian models(Model 4 and 5), revealed consistent and significant predictors of survival—specifically, 'Sex,' 'TotalLength,' and 'Weight.' Lower AIC scores in the improved BinomialModels 3 and Stepwise AIC Gaussian models (Models 5) indicated better model fits and are hence preferred for selection. These models also introduced important interaction terms, suggesting more complex relationships between predictors that could be essential for future investigations.

In terms of Generalized Linear Models (GLMs) with a binomial family, significant predictors such as 'Sex' with a positive coefficient indicate that being male increases the odds of survival, while 'TotalLength' and 'Weight' with negative coefficients suggest that increases in these variables are associated with lower odds of survival. The model fits were assessed using deviance, with a residual deviance of 130.21 on 125 degrees of freedom, and an AIC score of 152.21, suggesting the model's efficacy over a null model. These insights offer a robust foundation for further research and model optimization in predicting survival outcomes.

The reduced residual deviance compared to the null deviance and the relatively low AIC score suggest that this model provides a better fit than an intercept-only model.

The mathematical formula representing the log-odds of the outcome variable p (the probability of survival) can be expressed as follows (Model 3):

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) = {} & 15.8590 + 1.5958 \times \text{Sex(m)} \\ & - 0.4179 \times \text{TotalLength} \\ & + 0.6107 \times \text{BeakHead} \\ & - 0.8259 \times \text{Weight} \\ & + 51.4046 \times \text{Humerus} \\ & + 16.8185 \times \text{Sternum}\end{aligned}$$

In this formula, p is the probability of "Survival" (the dependent variable), and the other terms are the independent variables each multiplied by their respective coefficients from the model output.

Probabilities

$$p = \frac{e^{15.8590+1.5958\times\text{Sex(m)}-0.4179\times\text{TotalLength}+0.6107\times\text{BeakHead}-0.8259\times\text{Weight}+51.4046\times\text{Humerus}+16.8185\times\text{Sternum}}}{1+e^{15.8590+1.5958\times\text{Sex(m)}-0.4179\times\text{TotalLength}+0.6107\times\text{BeakHead}-0.8259\times\text{Weight}+51.4046\times\text{Humerus}+16.8185\times\text{Sternum}}}$$

And, e is the base of the natural logarithm, approximately 2.71828.

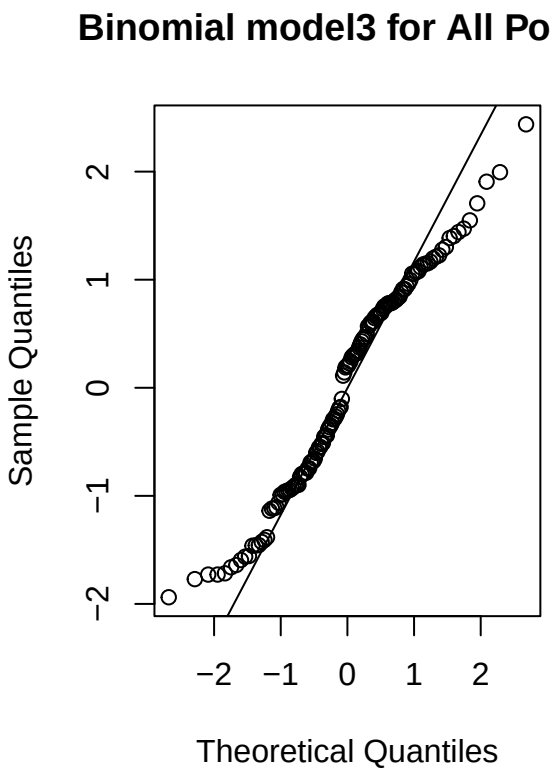
Discussion

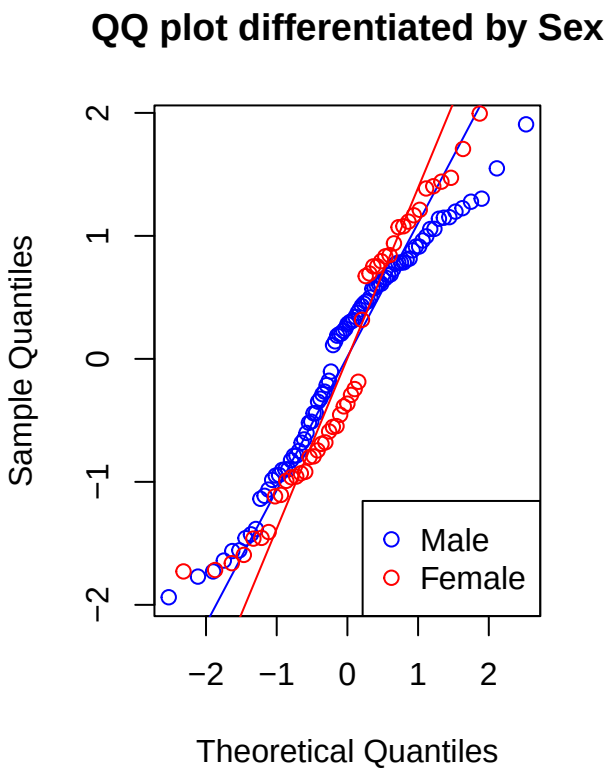
In the "Results" section, we presented the findings of our analysis of English sparrow survival data. The analysis revealed that wing length, total length and weight are significant predictors of survival. Birds with longer wings, small and low weight are more likely to survive. These findings could have ecological implications, as they suggest that certain morphological traits may confer advantages in terms of survival, possibly related to foraging or predator avoidance.

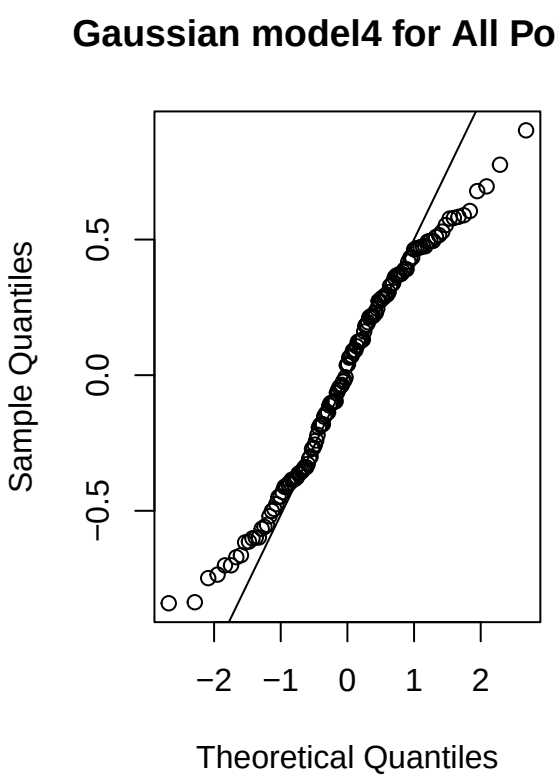
Observations:

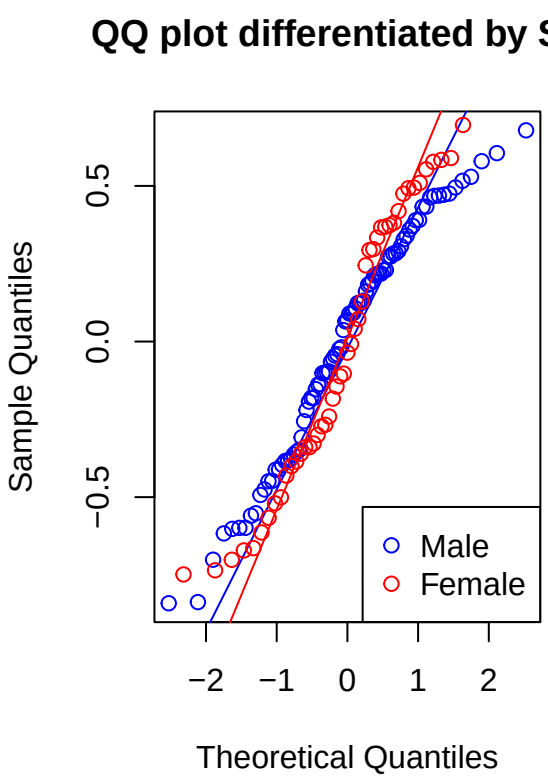
1. **Consistency Across Models**: 'Sex', 'TotalLength', and 'Weight' are consistently significant, which suggests these predictors are robustly associated with 'Survival'.
2. **Complexity vs Simplicity**: More complex models like Model 5 may fit the data better (as suggested by lower AIC) but are harder to interpret due to interaction terms, so not to overfit the data Model 4 works much better.
3. **Model Suitability**: Depending on the distribution of the 'Survival' variable which in our case is binary, Binomial model(model 3) is more appropriate than the others.
4. **Predictive Power**: Model 3 has the lowest AIC among the binomial models, and Model 5 among the Gaussian, implying these models might have better predictive power.
5. **Importance of Specific Measures**: For instance, the 'Humerus' variable was not significant in the linear or the initial binomial models but became significant when using stepAIC in the binomial model. This could imply that its significance was masked by other variables in the full model or that it's more appropriate for a non-linear relationship with the dependent variable.
6. **Check for Overfitting**: Complex models with many predictors and interaction terms are more likely to overfit the data. This realization is made since working on binomial, piossion and gaussian with quadratic and cubic terms gave the overfit model.

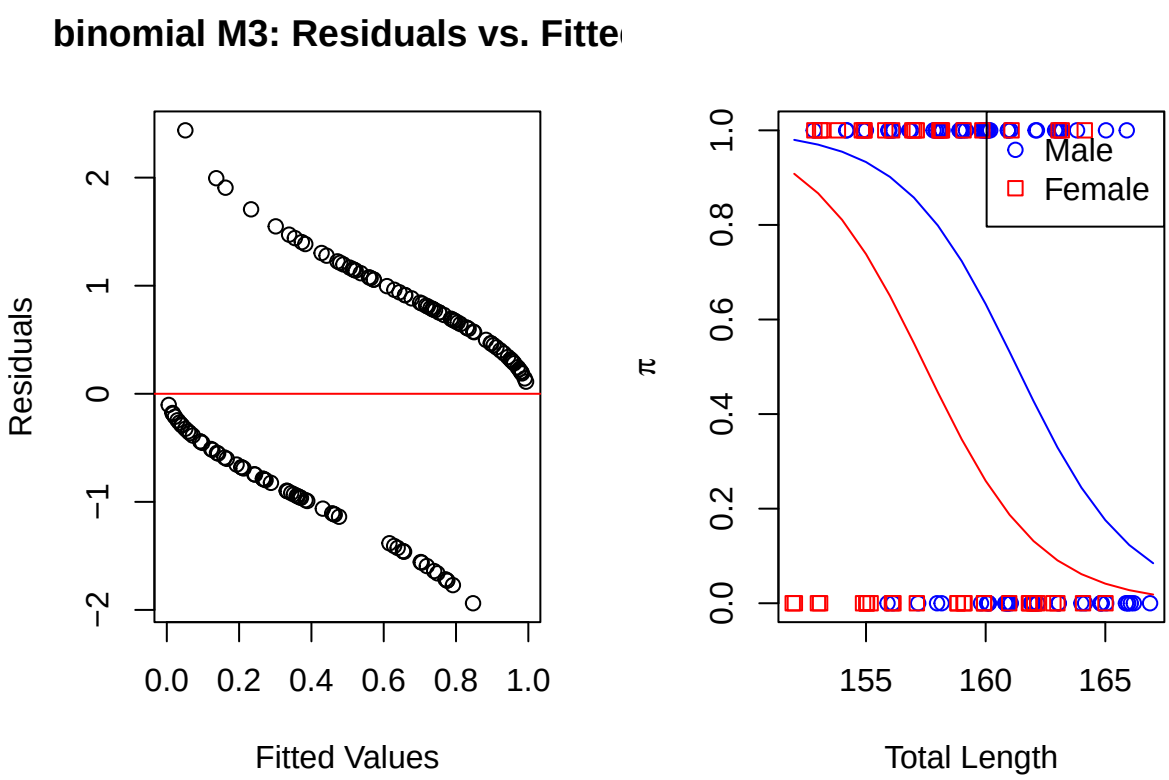
Given the nature of the dependent variable (Survival), binomial models are probably more appropriate.






Gaussian model4 for All Points
Sample Quantiles
0.5
0.0
−0.5
−2
−1
0
1
2
Theoretical Quantiles


QQ plot differentiated by Sex
Sample Quantiles
0.5
0.0
−0.5
−2
−1
0
1
2
Theoretical Quantiles
Male
Female



Limitations

It's essential to clarify that the statistical significance observed in this study indicates correlations rather than causative relationships. In other words, just because a variable is statistically significant doesn't mean it has a substantive impact on the outcome. Therefore, it's important to examine the magnitude of the coefficients to gauge the practical relevance of each predictor variable, in addition to their statistical significance.

While these results provide valuable insights, they should be viewed as a preliminary step requiring further validation. To comprehend the underlying mechanisms and ecological implications fully, additional research and more detailed ecological studies are indispensable. This study represents a simplified analysis, and a more exhaustive investigation would account for other factors, control variables, and ecological conditions to offer a well-rounded understanding of survival patterns in English sparrows.

References

1. BRUCE, H. PUGESEK. "The Bumpus house sparrow data: a reanalysis using structural equation models." Evolutionary Ecology 10 (1996): 387-404.
2. Johnston, Richard F., David M. Niles, and Sievert A. Rohwer. "HERMON BUMPUS AND NATURAL SELECTION IN THE HOUSE SPARROW PASSER DOMESTICUS." Evolution 26.1 (1972): 20-31.
3. BUTTEMER, WILLIAM A. "DIFFERENTIAL OVERNIGHT SURVIVAL BY BUMPUS'HOUSE SPARROWS: AN ALTERNATE INTERPRETATION'." The Condor 94.4 (1992): 944-954.
4. Bentler, P. M., and Douglas G. Bonett. "Significance Tests and Goodness of Fit in the Analysis of Covariance Structures." Psychological Bulletin 88.3 (1980): 5SS-606.
5. BROWNE, MW. "Asymptotically distribution-free methods for the analysis of covariance structures." British Journal of Mathematical and Statistical Psychology 37 (1984): 62-83.

6. Grant, Peter R. “Centripetal selection and the house sparrow.” Systematic Biology 21.1 (1972): 23-30.

7. Harris, J. Arthur. “A neglected paper on natural selection in the English sparrow.” The American Naturalist 45.533 (1911): 314-318.

8. O’Donald, Peter. “A FURTHER ANALYSIS OF BUMPUS’DATA: THE INTENSITY OF NATURAL SELECTION.” Evolution 27.3 (1973): 398-404.

9. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2023.