Morphological Determinants of Sparrow Survival: An Analytical Exploration of the Bumpus Dataset

Angela Jacinto 23778435

Executive summary

This report focuses on modelling the survival probabilities of sparrows, aiming to identify if certain morphological traits or characteristics correlate with higher or lower survival rates. Additionally, it seeks to determine the simplest model that effectively demonstrates this relationship.

The provided "Bumpus" dataset comprises measurements and characteristics of sparrows collected during a specific event in history. On February 1, 1898, during a winter storm in Providence, 136 sparrows were brought to Professor Hermon C. Bumpus's anatomy lab. Of these sparrows, 72 were successfully revived, while 64 did not survive. The data includes various morphological traits and characteristics, including sex, survival, total length, alar extent, weight, beak head measurements, humerus, femur, tibiotarsus, skull width, and sternum measurements. The logistic regression model, which includes only main effects, emerged as the most effective model for predicting the survival probabilities of sparrows. This model, characterized by its simplicity and absence of interaction terms, demonstrated a preferable fit compared to the more complex model, which incorporated a few interaction terms. Including interaction terms did not show a statistically significant improvement in the model's fit. The chosen model revealed that the total length and weight of a sparrow, as well as being female, are negatively associated with the log odds of survival. Conversely, the size of the Humerus and Sternum displayed a positive association with the log odds of survival in sparrows.

Introduction

The context of this study revolves around the severe winter storm that occurred on February 1, 1898 where 136 sparrows were brought in by Professor Hermon C. Bumpus. Of these sparrows, 52.94% survived while the remaining 47.06% died. The dataset provides comprehensive measurements and attributes of these sparrows, which was analyzed to determine which traits most influenced their chances of survival during the storm. Moreover, the goal of this study is to identify patterns that highlight the key factors affecting survival in such conditions.

After extensive analysis, one aspect of the study demonstrates the survival rate and gender disparity. Out of 136 sparrows, a little over half survived the storm. There is a gender disparity in overall survival rates, with a higher proportion of male sparrows surviving. Several morphological measurements and characteristics were found to be significant influences on survival. Multiple models were constructed but mainly focused on a model that had the best balance between simplicity and interpretability as it is just as effective in predicting sparrow survival compared to complex models that include interaction terms based on key metrics such as AIC score and residual deviance.

The Pearson residuals analysis of the selected model showed a satisfactory fit, with residuals fairly distributed around the zero line, indicating that the model's predictions align well with the observed outcomes. The residuals suggest potential heteroscedasticity, which might be a concern for model assumptions. Plots like "Total Length vs. Survival" and "Humerus vs. Survival" visually reinforced the findings from the models. For example, as sparrows get larger in terms of total length, their survival probability decreases, especially for females. However, an increase in humerus size is associated with a higher survival likelihood, more pronounced for males compared to females.

In summary, the study successfully identified several key morphological measurements and characteristics that influence the survival of sparrows during severe winter storms, with evident gender disparities in the survival rates. The constructed models, especially the simpler ones, were effective in capturing these relationships and predicting survival outcomes.

Methodology

We first explore the data through numerical and graphical summaries. During the preprocessing phase, the sex of sparrows was encoded numerically, where females were labeled as '1' and males as '0'. Further using males as the data's reference level. Furthermore, the survival variable was converted to a numeric format for subsequent analyses.

To discern how the survival of sparrows is influenced by the various measurements and characteristics, a logistic regression model was fitted to the data with the log odds of survival as the response. The process involved multiple strategies for model building, such as forward and backward selection techniques. A comprehensive model which incorporated two-way interactions of the main effects was used. The "stepAIC" function was then used to perform both forward and backward selection to refine the model. This process added and removed predictors, guided by their impact on the model's Akaike Information Criterion (AIC) and residual deviance values. Alongside these models, a simpler model where only main effects were considered was also tested. The intention behind this simpler model was to determine if it could achieve a similar level of predictive accuracy while being more straightforward to interpret.

Several models were obtained from this process, each with different combinations of main effects and interaction terms. To ensure that the findings in the models are clear and robust, the models were examined based on both automated (stepAIC) and manual selection methods. The primary goal was to identify a model that has a balance between simplicity and a good fit in the model, ensuring that the selected model could accurately predict sparrow survival based on morphological measurements and other characteristics without the model being too complex to interpret.

Once the final model was chosen, further diagnostic procedures were conducted to assess its validity and reliability. Pearson residuals analysis was performed to evaluate the model's goodness of fit. This analytical technique offered insights into potential violations of the logistic regression assumptions, such as linearity, independence, homoscedasticity, and normality of residuals. Calculating the odds of survival given each predictor was also done to provide a more tangible interpretation of how changes in each predictor influence the survival probability. This approach translated the log-odds coefficients into a more intuitive probability scale, thereby enhancing the understanding of the model's outcomes.

Lastly, a drop-in deviance test was conducted to compare our chosen model against alternative models. This test helped determine if the inclusion of certain predictors significantly enhanced the model's fit. Additionally, confidence intervals for each predictor were calculated to quantify the uncertainty around each coefficient estimate, offering a range within which the true value of the effect size is likely to lie. These intervals served as a crucial tool for assessing the precision and reliability of the model's predictions.

Results

	Died	Survived	
Male	36	51	87
Female	28	21	49
	64	72	136

Table 1: Contingency table

We first examined the categorical predictors from the data to discern the meaning of the values and variables that will be dealt with. The contingency table illustrates a higher representation of males compared to females. There are a total of 136 sparrows in the dataset where 87 are males, constituting approximately 64% of the dataset. The remaining 49 female sparrows make up about 36% of the dataset. Among the males, a greater proportion survived while a greater proportion of the females died, which illustrates the higher survival rate of males compared to females.

Variable	Description	Summary
Sex		Male (0) = 87, Female (1) = 49

Total Length	measured from tip of the beak to the tip of the tail (mm)	min: 152, median: 160, max: 167, sd: 3.5608
Alar Extent	measured from tip to tip of the extended wings (mm)	min: 230, median: 246, max: 256, sd: 5.5210
Weight	weight of the bird (g)	min: 22.6, median: 25.55, max: 31, sd: 1.4752
Beakhead	length of beak and head, measured from tip of the beak to the occiput (mm)	min: 29.8, median: 31.6, max: 33.4, sd: 0.7024
Humerus	length of humerus (inches)	min: 0.659, median: 0.733, max: 0.78, sd: 0.0231
Femur	length of femur (inches)	min: 0.653, median: 0.713, max: 0.767, sd: 0.0241
Tibiotarsus	length of tibiotarsus (inches)	min: 1.011, median: 1.133, max: 1.230, sd: 0.0407
Skull Width	width of skull measured from the postorbital bone of one side to the postorbital bone of the other (inches)	min: 0.551, median: 0.602, max: 0.64, sd: 0.0150
Sternum	length of keel of sternum (inches)	min: 0.734, median: 0.841, max: 0.927, sd: 0.0.396

Table II: Summary of data

The standard deviations in the table above provide an insight into the range of physical characteristics among the sparrows. These variations may influence the relationship between each predictor and the probability of survival. The predictor Alar Extent shows the highest standard deviation, indicating significant variability in the wing size of sparrows. A predictor with higher variability may contribute to an uncertain understanding of its relationship with the outcome since it offers a wide range of data.

$$logit(P(Survival)) = 17.4548 - 0.4179 \times TotalLength + 51.4046 \times Humerus - 1.5958 \times Sex - 0.8259 \times Weight + 16.8185 \times Sternum + 0.6107 \times BeakHead$$

Model Equation I: “currentmodel”

Model equation I presents a model constructed using the stepAIC function using forward selection. This model, characterized by its significantly simpler equation, illustrates the best balance between simplicity and interpretability among all the models considered. Despite its simplicity, the model exhibits a strong fit, with an AIC score of 145.57 and a residual deviance of 131.57.

$$logit(P(Survival)) = 26.0266 - 20.9003 \times Sex - 0.6505 \times TotalLength + 0.1351 \times AlarExtent - 0.8358 \times Weight + 0.6100 \times Beakhead + 46.0436 \times Humerus + 16.3525 \times Sternum + 0.4221 \times Sex{:}TotalLength - 0.1956 \times Sex{:}AlarExtent$$

Model Equation II: “m3”

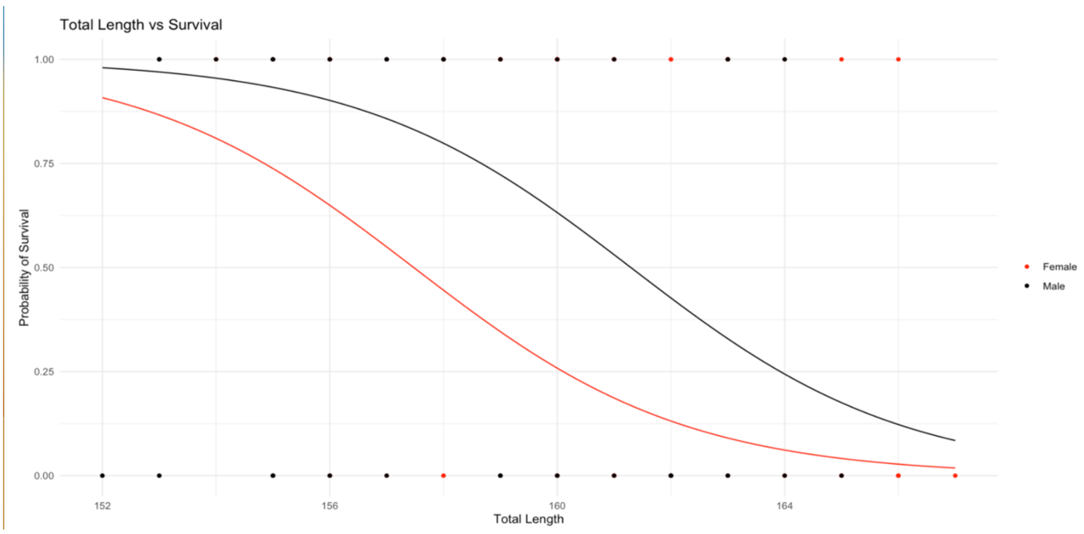
Model equation II presents a model obtained through manual forward selection, aiming for simplicity while maintaining a reasonable fit. It shows an AIC of 146.81 and a residual deviance of 126.81, slightly higher than the previous model (Model Equation I). This model includes main effects and two interaction terms: "Sex:TotalLength" and "Sex:AlarExtent." The predictors, "TotalLength," "Weight," "Humerus," and "Sex:TotalLength" are statistically significant.

	Residual Df	Residual Deviance	Df	Deviance	Pr(>Chi)

Model Equation I	129	131.57			
Model Equation II	126	126.81	3	4.7593	0.1903

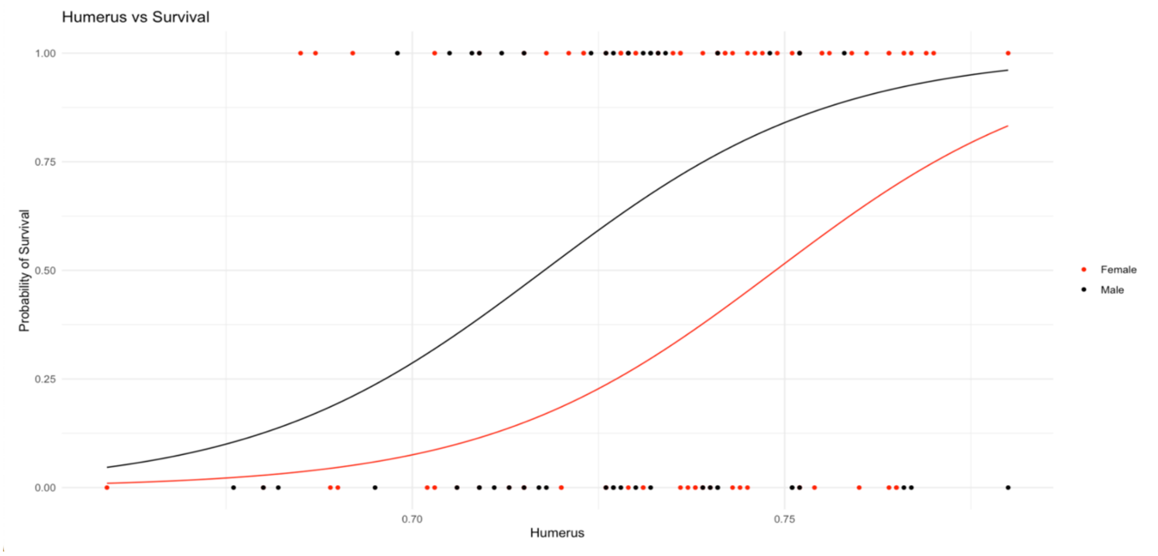
Table III: Drop-in-Deviance/ Goodness-of-Fit test

Given this, model equation I, "currentmodel," is preferred due to its optimal balance between simplicity and interpretability, alongside its strong statistical fit. This model was constructed using the stepAIC function with forward selection, which is a methodical approach aimed at enhancing model efficiency by including only the most significant predictors. Its significantly simpler equation contributes to its ease of interpretation without compromising the model's predictive capability. Despite its relative simplicity, "currentmodel" demonstrates a strong fit to the data, as evidenced by its lower Akaike Information Criterion (AIC) score of 145.57 and a residual deviance of 131.57. Lower AIC values typically suggest a better model fit with fewer unnecessary parameters, and in this case, the AIC of "currentmodel" is slightly lower than that of "m3" (146.81). A lower residual deviance also points towards a better fit of the model to the observed data. In contrast, Model Equation II, "m3," though also aiming for simplicity, incorporates additional complexity with the inclusion of interaction terms, "Sex:TotalLength" and "Sex:AlarExtent". While these interactions offer insights into the combined effects of variables, they add complexity to the model, potentially reducing its interpretability. Moreover, the slightly higher AIC and residual deviance of "m3" suggest a less efficient model fit compared to "currentmodel." Therefore, given the goals of achieving a balance between simplicity and accurate fit, "currentmodel" is the preferred choice. It effectively captures the essential relationships in the data with fewer parameters and a more straightforward interpretative framework, making it a more suitable model for understanding the survival chances of sparrows in relation to their morphological traits. Furthermore, a chi-square test to compare the fits of the two models indicates that the more complex model "m3" does not significantly improve the fit over "currentmodel". With a p-value of 0.1903, the test does not support a statistically significant difference in the fit of the two models at the 0.05 significance level. This suggests that the interaction terms added in "m3" do not provide a statistically significant enhancement in explaining the survival of sparrows. Consequently, "currentmodel" is selected for its sufficiency in capturing the predictive relationships without the additional complexity of interaction terms.



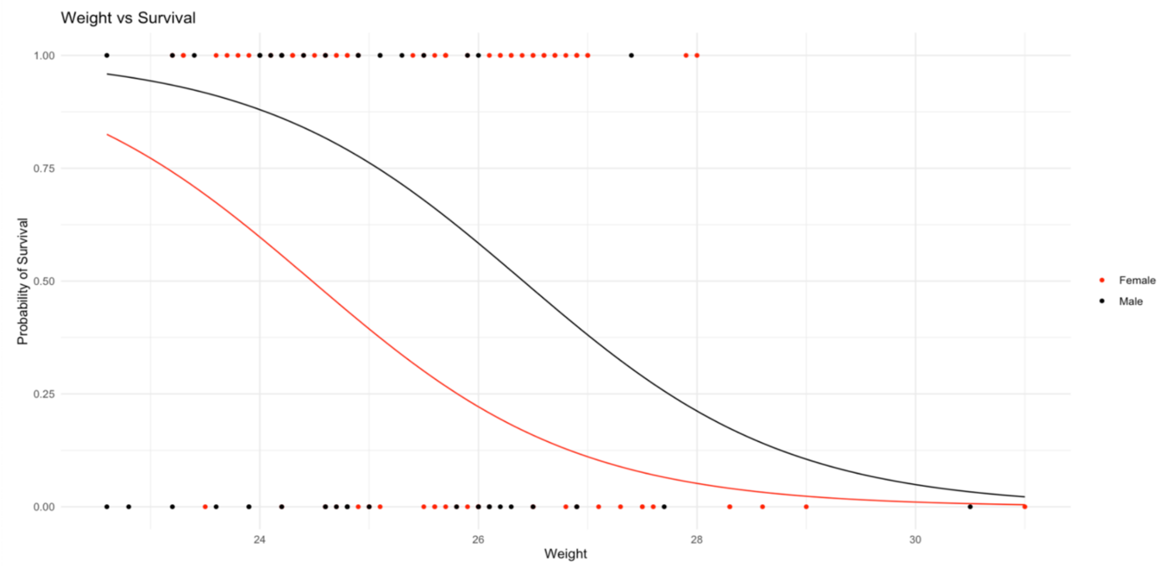
Plot I: Total Length vs Probability of Survival

Plot I visualizes the predicted probability of sparrow survival based on the predictor total length separately for male and female sparrows. It uses the chosen logistic regression model to predict survival probabilities over the range of the variable total length, holding other variables at their mean values. It illustrates how changes in the total length of a sparrow affect the predicted probability of survival by sex. The plot above indicates that as a sparrow increases in total length, their predicted survival probability decreases. However, at any given length, male sparrows generally have a higher predicted survival chance than females. The lines for each gender appear to deviate as sparrows get larger, suggesting an increasing difference in survival rates between the genders with size.



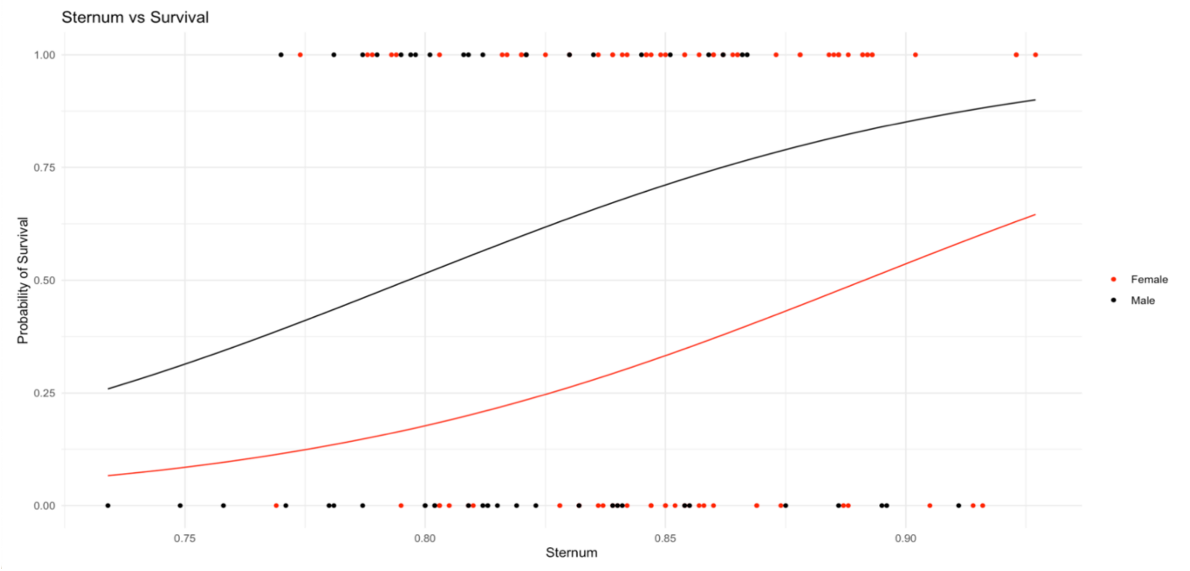
Plot II: Humerus vs Probability of Survival

Plot II shows that sparrows of both genders with shorter humerus lengths predominantly show zero survival probability. This trend shifts just below the average humerus size, where survival rates slowly increase. As humerus size increases, male sparrows consistently exhibit a higher survival likelihood than females. The difference in survival rates between male and female sparrows becomes more pronounced as the humerus size grows.



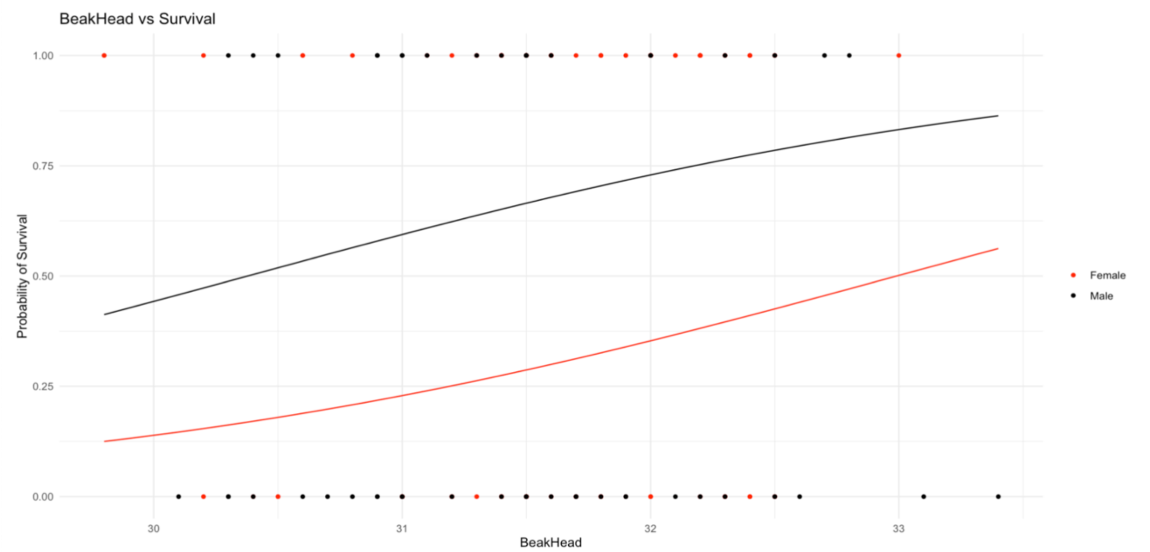
Plot III: Weight vs Probability of Survival

Plot III represents a noticeable trend; a higher proportion of female sparrows have a survival probability of one. As weight increases, the likelihood of survival for sparrows decreases, especially evident when reaching a weight of 28, where survival probability approaches zero. It is worth noting that the average weight in the dataset is 25.58. The male survival curve lies consistently above the female curve. However, both demonstrate a pronounced deviation in survival probabilities as weight increases.



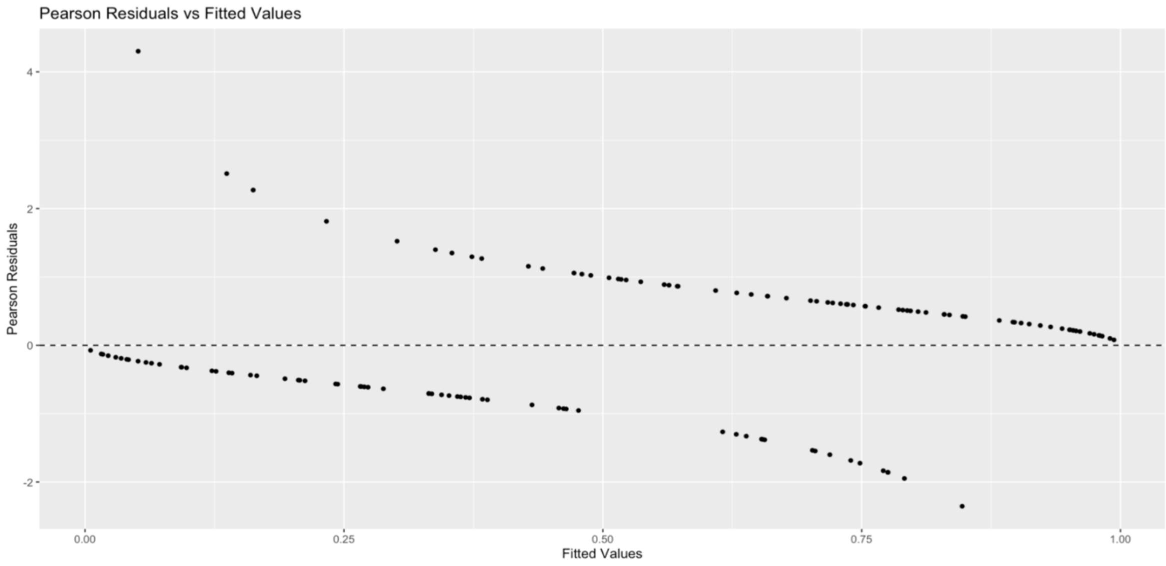
Plot IV: Sternum vs Probability of Survival

Plot IV suggests that the data shows a clear distinction between male and female sparrows. As sternum length increases, male sparrows demonstrate a rise in survival probability, nearing certainty for longer sterna. In contrast, female sparrows have a more moderate increase in survival chances with sternum length, between 0.50 and 0.75 at the maximum length of the dataset. The deviation between the regression lines suggests that a longer sternum offers a more significant survival advantage for males compared to females.



Plot V: Beakhead vs Probability of Survival

Plot V shows that the regression line increase in beakhead length is associated with a gradual rise in survival probability for both genders. Males demonstrate a survival probability shift from around 0.40 at smaller beakhead size to approximately 0.88 at beakhead size above 33. In contrast, females begin at a lower survival probability near 0.12 and only reach just below 0.62, even with the longest beakhead measurements. This deviation suggests that as beakhead size grows, the survival advantage becomes more distinct between males and females.



Plot VI: Pearson Residuals vs Fitted Values

Pearson residuals were used to assess the chosen model's goodness of fit. The residuals seem to be fairly distributed around the zero line, and there doesn't seem to be any pattern suggesting a departure from normality, which suggests that the model's predictions align well with the observed outcomes and indicate a satisfactory model fit. While the plot does not have any pronounced U-shapes or curves, there is a subtle funnel shape towards the higher end of fitted values, which indicates potential heteroscedasticity. Furthermore, there is an uneven spread of residuals— evidence of heteroscedasticity. This tells us that the model's precision varies across the fitted values. The plot also shows a clear pattern of the residuals with points forming a funnel shape as the fitted values increase, which suggests that the model is not entirely capturing the relationship between the dependent and independent variables.

Odds of Survival (Lower Quartile)	Odds of Survival (Upper Quartile)	Odds Ratio (Lower vs Upper)	Percentage Change	Multiplicative Change	Confidence Interval
9.858036e-21	1.868166e-23	0.001895069	-99.81049 %	0.6584281	-0.616644, -0.219156

Table IV: Odds of Survival Given Total Length

Analysis of the survival odds across quartiles of Total Length reveals that sparrows in the lower quartile have very low odds of survival at 9.858036e-21, and this decreases further in the upper quartile to an even lower odds of 1.868166e-23. The computed odds ratio of 0.001895069 confirms that the survival odds at the upper quartile are substantially lower than those at the lower quartile, with a dramatic percentage change of -99.81049%, illustrating a decrease in survival odds with increasing Total Length. Additionally, the odds of survival diminish by about 34% with each unit increase in Total Length, as indicated by the value of 0.6584281.

A confidence interval analysis is utilized to estimate the range in which the true value of the regression coefficient for the predictor "Total Length" likely falls, with a 95% confidence level. The calculated interval, ranging from -0.616644 to -0.219156, suggests a negative association between the predictor and the log odds of survival for sparrows; this is because the entire interval lies below zero. The exclusion of zero from this interval further reinforces the statistical significance of the predictor's effect, confirming its significant impact on survival at the 5% significance level.

Odds of Survival (Lower Quartile)	Odds of Survival (Upper Quartile)	Odds Ratio (Lower vs Upper)	Percentage Change	Multiplicative Change	Confidence Interval
1.961206e+22	9.858616e+24	502.6814	50168.14 %	2.112196e+22	22.33819, 80.47101

Table V: Odds of Survival Given Humerus

When examining survival odds across quartiles for the predictor Humerus, it is found that sparrows in the lower quartile have exceedingly high odds of survival at 1.961206e+22, and this likelihood of survival climbs even more in the upper quartile, reaching 9.858616e+24. The odds ratio calculated at 502.6814 suggests that sparrows at the upper quartile have survival odds approximately 503 times higher than those at the lower quartile, with the percentage change of 50168.14% marking a dramatic rise in survival odds with increasing Humerus size. Furthermore, the value of 2.112196e+22 underscores the magnitude of this effect, suggesting that each unit increase in Humerus size multiplies the odds of survival by a factor of 2.112196e+22.

The confidence interval for the coefficient of Humerus, ranging from 22.33819 to 80.47101, suggests a positive relationship between the size of the humerus and the log-odds of sparrow survival, despite indicating some uncertainty about the exact magnitude of this effect. The interval lies above zero, reinforcing the significance of the association—an increase in humerus size is correlated with an increase in the log-odds of survival. Moreover, the absence of zero from this confidence interval affirms the statistical significance of the predictor's effect on survival at the 5% level.

Odds of Survival (Lower Quartile)	Odds of Survival (Upper Quartile)	Odds Ratio (Lower vs Upper)	Percentage Change	Multiplicative Change	Confidence Interval
7717492	38064780	4.932273	393.2273 %	0.2027463	-2.631072 - 0.560528

Table VI: Odds of Survival Given Sex

The odds of survival calculated for females stand at 7717492, which are overshadowed by the even higher odds for males, calculated to be 38064780. This substantial discrepancy is further quantified by an odds ratio of 4.932273, indicating that males have approximately 4.93 times higher odds of survival than females. The percentage change in odds of 393.2273% underscores the substantial disparity between the sexes. Furthermore, the value of 0.2027463 conveys that being female is associated with an approximate 80% decrease in the odds of survival when compared to their male counterparts.

The confidence interval for the coefficient of sex, extending from -2.631072 to -0.560528, further reinforces a negative relationship between being a female sparrow and the log odds of survival; this is evidenced by the entire interval falling below zero. This interval indicates that an increase in the value of the variable sex, which represents a change from male (0) to female (1), is associated with a decrease in survival odds. Furthermore, the absence of zero within this confidence interval points to the statistical significance of the variable's impact on survival, confirming its importance at the 5% significance level.

Odds of Survival (Lower Quartile)	Odds of Survival (Upper Quartile)	Odds Ratio (Lower vs Upper)	Percentage Change	Multiplicative Change	Confidence Interval
0.2980363	0.0002892845	0.0009706351	-99.90294 %	0.4378408	-1.286108 - 0.365692

Table VII: Odds of Survival Given Weight

When examining the odds of survival across the weight spectrum, sparrows in the lower quartile of weight exhibit relatively higher survival odds compared to those in the upper quartile, where the odds drop significantly to 0.0002892845. This contrast is further quantified by an odds ratio of 0.0009706351, indicating that the survival odds at the upper quartile of weight are considerably lower than those at the lower quartile, with a percentage change of -99.90294%. Moreover, the odds of survival are halved for every increase in weight, decreasing by approximately 57% for each unit increase, underscoring the substantial impact of weight on the likelihood of sparrow survival.

The confidence interval for the coefficient of weight, ranging from -1.286108 to -0.365692, underscores a negative relationship between weight and the log-odds of survival. The entire interval lies below zero, which implies that an increase in weight corresponds to a decrease in the log-odds of survival. The absence of zero in this interval is crucial, as it signifies that the influence of weight on survival is not only negative but also statistically significant at the 5% level.

Odds of Survival (Lower Quartile)	Odds of Survival (Upper Quartile)	Odds Ratio (Lower vs Upper)	Percentage Change	Multiplicative Change	Confidence Interval
8.745663e+12	2.246466e+14	25.68663	2468.663 %	20145671	0.784132 32.852868

Table VIII: Odds of Survival Given Sternum

When focusing on survival odds across different quartiles of the variable sternum, the lower quartile yields remarkably high odds of survival at 8.745663e+12, and this increases even more for the upper quartile, reaching 2.246466e+14. This increase is further quantified by an odds ratio of 25.68663, suggesting that the survival odds at the upper quartile of sternum are about 25.69 times higher than those at the lower quartile. The astounding percentage change of 2468.663% between these quartiles underscores a significant increase in survival odds with an increase in sternum size. Moreover, the extremely large outcome of “exp(16.8185)” at 20145671, demonstrates that even a small increase in the sternum size of a sparrow can lead to a substantial increase in the odds of survival, reflecting the considerable impact of sternum size on a sparrow’s odds of survival.

The confidence interval for the coefficient of sternum, spanning from 0.784132 to 32.852868, effectively reinforces the positive relationship between the size of a sparrow's sternum and the log-odds of survival. This interval, situated entirely above zero, clearly suggests that an increase in sternum size correlates with a corresponding increase in the log-odds of survival. As zero is not included within this range, it is indicative of the statistical significance of the sternum's impact on survival, confirming its importance at the 5% significance level. This finding highlights the substantial influence that sternum size has on the likelihood of survival in sparrows.

Conclusion and Discussion

Of the 136 sparrows analyzed, approximately 53% survived the storm, with a higher survival rate among males than females. Through the chosen logistic regression model, several morphological measurements and characteristics were identified as influential predictors of survival. The variable Humerus revealed a substantial positive relationship with the survival of sparrows: a one-unit increase in humerus size is associated with an increase of 51.4046 in the log-odds of survival, indicating a strong positive association. Similarly, the variable Sternum also revealed a positive relationship: a unit increase in the size of a sparrow's sternum is associated with an increase of 16.8185 in the log-odds of survival, indicating a favorable association. On the other hand, a negative coefficient of -0.4179 for the variable total length implies that an increase of one unit in total length is associated with a decrease in the log-odds of survival. The model also revealed a negative association between a sparrow's weight and its odds of survival; specifically, for each unit increase in weight, the log-odds of survival decrease by 0.8259. Furthermore, gender differences played a significant role, with male sparrows generally demonstrating a higher probability of survival across various morphological measurements. In this analysis, where males serve as the baseline group and are coded as 0, the coefficient for Sex reflects the difference in log-odds of survival for female sparrows. Specifically, the coefficient of -1.5958 indicates that female sparrows have lower log-odds of survival compared to males, revealing a negative association between being female and the likelihood of survival. The size of the beakhead was not statistically significant at $\alpha = 0.05$ (5%). The plots further proved that sparrows with longer total lengths, especially females, showed decreased survival probabilities. Sparrows with longer humeri had increased survival odds, with males having a significantly higher probability of survival than females. A longer sternum also offered a more distinct survival advantage for males compared to females. An increase in weight reduced survival chances for both genders, but the decline was more pronounced for females. Lastly, the beakhead size increase led to a gradual rise in survival probability, with males having a significant advantage over females.

The Pearson residuals analysis demonstrated a satisfactory model fit, suggesting that the model's predictions align closely with observed outcomes. However, potential heteroscedasticity was observed, which might affect the reliability of predictions for sparrows with specific characteristics.

There are potential issues and limitations to the study that can be potentially addressed. There was an observed heteroscedasticity in the Pearson residuals plot, which suggests that the model's assumption might not be fully met. As a result, this can affect the accuracy of the model's predictions. On the other hand, a broader data scope could be considered in the future as the dataset was specific to sparrows that were affected by a certain environmental condition. Other analytical techniques such as support vector machines or machine learning models can potentially offer better predictive accuracy, especially with dealing with complex non-linear relationships within the data.

References

American Ornithological Society. (2018). Professor Bumpus and His Sparrows. Retrieved [2/10/23], from https://americanornithology.org/professor-bumpus-and-his-sparrows/