Predicting Body Mass of Antarctic Penguins using Logistic Regression

By: Bishwadeep Bhattacharyya, Anum Damani, Zeyan Huang, Yifan Shen, and Yuxin Zhang

STATS 402

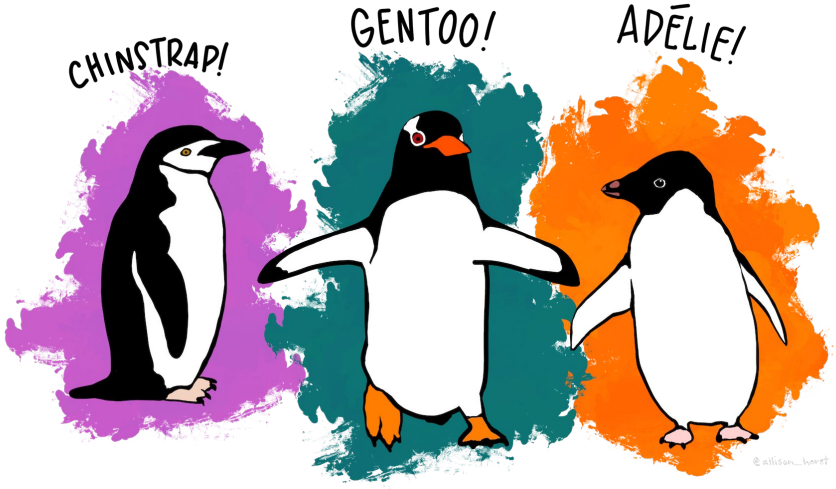
Fall 2023

Abstract

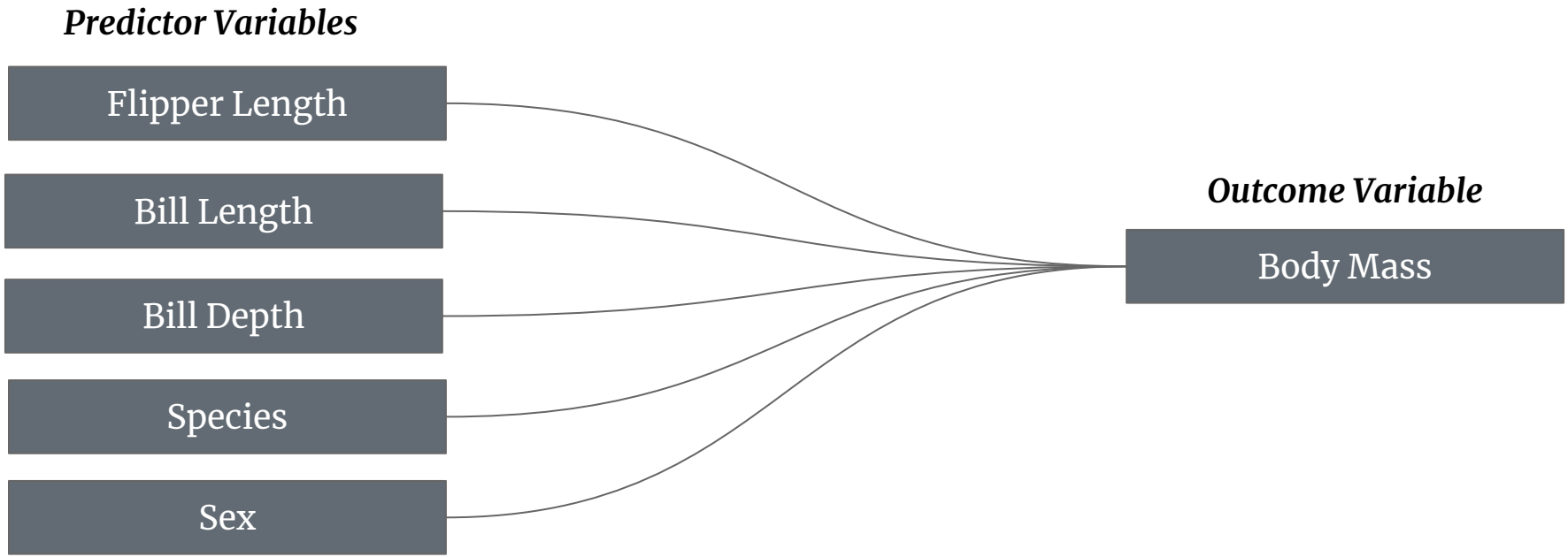
More than half of the world's penguin species are endangered.[1] Some of the reasons why penguins are vulnerable are habitat loss, pollution, disease, reduced food availability, and invasive predators. Weight can be a good indicator of health. Although higher body weight can make a penguin more vulnerable to predators, it can be beneficial in terms of breeding.[2] We were interested in determining whether body mass of Antarctic penguins can be predicted from their characteristics, specifically species, bill length, bill depth, flipper length, and sex. Using logistic regression, we found that these predictor variables can be used to predict whether penguins fall into the "above median body mass" category with high accuracy (**87%**). To improve our prediction accuracy, we can consider exploring more interaction terms, finding a larger dataset, and using multinomial regression.

Research Question

Can the body mass of Antarctic penguins be predicted from species, bill length, bill depth, flipper length, and sex?



Roadmap


Predictor Variables
Flipper Length
Bill Length
Bill Depth
Species
Sex
Outcome Variable
Body Mass

Exploratory Data Analysis (EDA): Numerical Variables

- Prepared for EDA by dropping all missing values in the data set[3]
  - After doing so, resulting dataset contained 333 observations
- Created histograms and boxplots (using symbox package) and determined that transformations are not necessary
- Created correlation matrix of numerical variables:
  - Flipper length and bill length are moderately, positively correlated
  - Flipper length and bill depth are moderately, negatively correlated

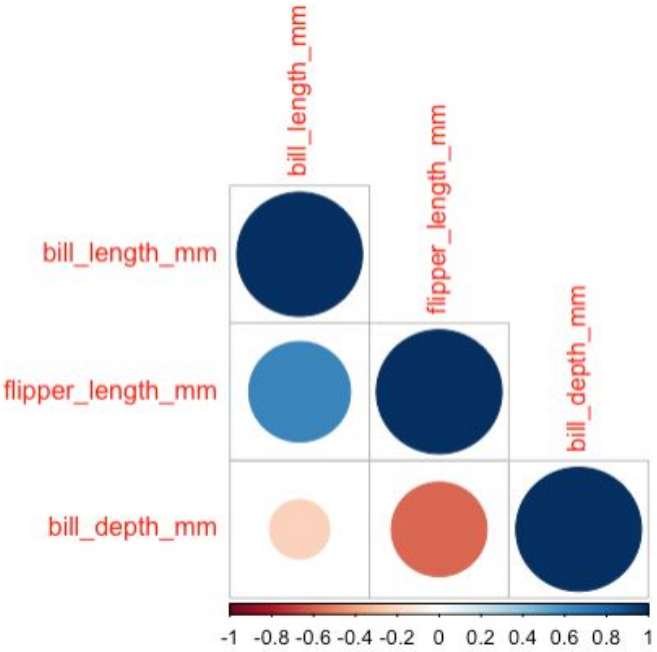


Figure 1: Heat Map for Numerical Variables

EDA: Categorical Variables

- ‘body mass’ variable (numerical) did not need transformations, either
  - We changed it to a categorical by cutting it at the median
- Checked the levels of the three categorical variables:
  - Body Mass: 2 levels
  - Species: 3 levels
  - Sex: 2 levels
- Created a frequency table for body mass, species, and sex

Body Mass	
Below Median	Above Median
171	161
Species		
Adelie	Chinstrap	Gentoo
146	68	119
Sex	
Female	Male
165	168

Figure 2: Table of Frequencies for Categorical Variables

EDA: Interaction Effect

- We were interested in determining the effect of species on bill length so we ran ANOVA
  - There is a statistically significant difference in group means at any level of ‘species’
- The interaction effect plot (Figure 3) shows the combined effect of species and bill length
  - As bill length increases, body mass tends to increase for each of the three species
  - It is clear that, as bill length increases, body mass tends to be higher for Gentoo penguins in comparison to Adelie and Chinstrap penguins

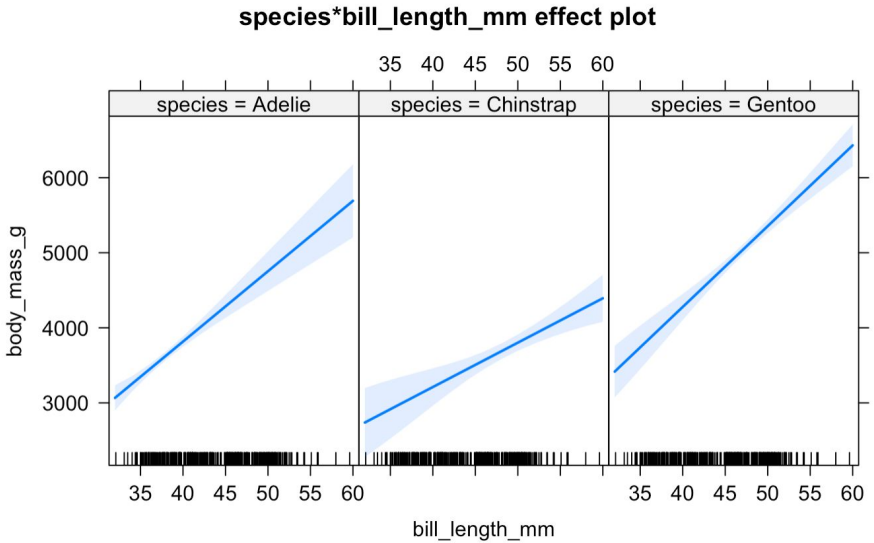


Figure 3: Plot of Interaction Effect between Species and Bill Length

EDA: Multiple Linear Regression

- Fitted a MLR model to determine the effect of species, bill length, flipper length, bill depth, sex, and the combined effect of species and bill length on body mass
  - Combined effect of species and bill length was not statistically significant so it was removed from the model
- Fitted another MLR model without this interaction term
  - Findings:
    - Species, bill length, flipper length, bill depth, and sex are all statistically significant
    - Multiple R-Squared: 0.875
    - Adjusted R-Squared: 0.8727

Logistic Regression Model

$log(p/(1-p))$

$= logit(p) = -36.467$

$- 3.120 \cdot$ speciesChinstrap

$+ 6.441 \cdot$ speciesGentoo

$+ 0.167 \cdot$ bill_length_mm

$+ 0.391 \cdot$ bill_depth_mm

$+ 0.099 \cdot$ flipper_length_mm

$+ 2.803 \cdot$ sexmale

Note: p = P{Y=1} = odds in favor of "success" (where "success" is a body mass that is higher than the median)

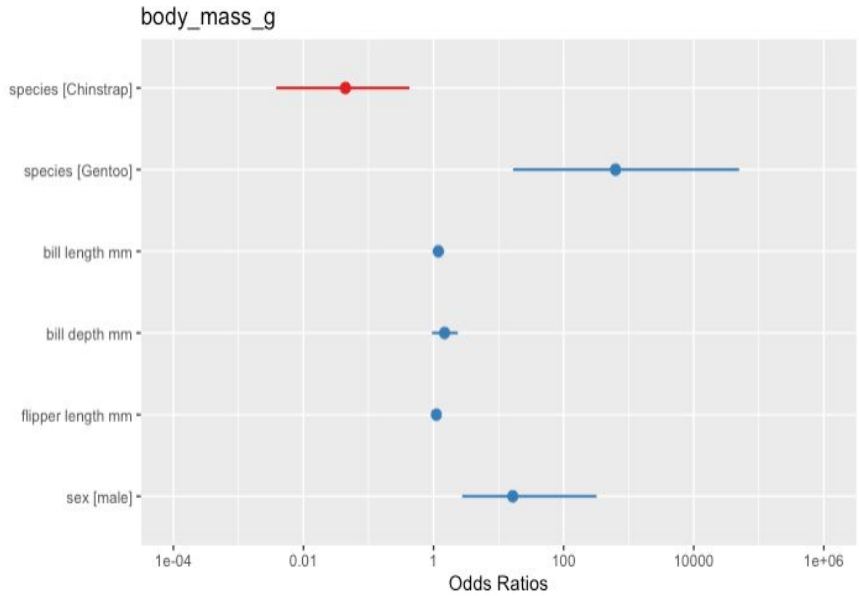


Figure 4: Plot of Odds Ratios

Odds Ratios, 95% CIs, & P-Values

Predictor	Odds Ratio	95% CI of Odds Ratio	p-value
speciesChinstrap	0.04417909	(0.0038, 0.426)	0.00903 **
speciesGentoo	627.3212	(16.693, 49903.100)	0.00116 **
bill_length_mm	1.182315	(0.970, 1.459)	0.10431
bill_depth_mm	1.479180	(0.945, 2.369)	0.09239
flipper_length_mm	1.104179	(1.028, 1.194)	0.00867 **
sexmale	16.50186	(2.750, 320.770)	0.01106 *

Figure 5: Table of Predictors, Odds Ratios, 95% CIs of Odds Ratios, and P-Values

Goodness of Fit for Log Reg Model

- Pearson's Chi-Squared Test
  - Null hypothesis: The logistic regression model is a good fit for the data
  - The calculated value of Pearson's Chi-Square (303.725) is less than the critical value of Chi-Square (368.0416) based on model's residual degrees of freedom
    - Test Result:
      - Fail to reject the null hypothesis
      - Suggests that the logistic regression model is a good fit for the data

Goodness of Fit for Log Reg Model

- Marginal Model Plots:
  - The curves based on the data are similar to the curves based on the model. In other words, the fitted values are similar to the observed data, indicating that the model is a good fit for the data.

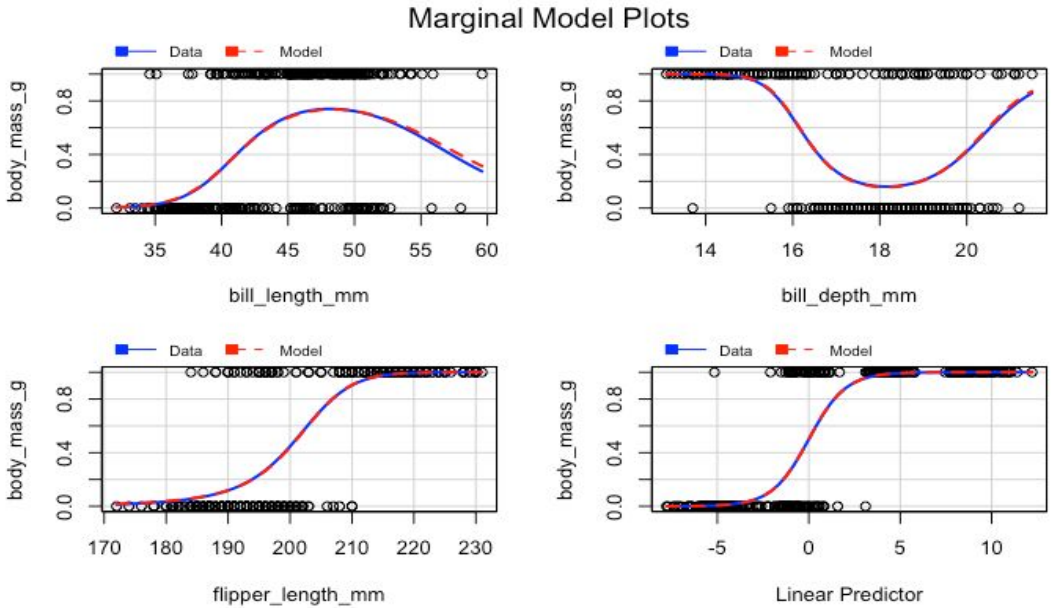


Figure 6: Marginal Model Plots

Goodness of Fit for Log Reg Model

- Residual Plot:
  - The plot shows that observations 10, 186, 274, and 313 are potentially concerning. Some of these points are high leverage but they are not outliers.

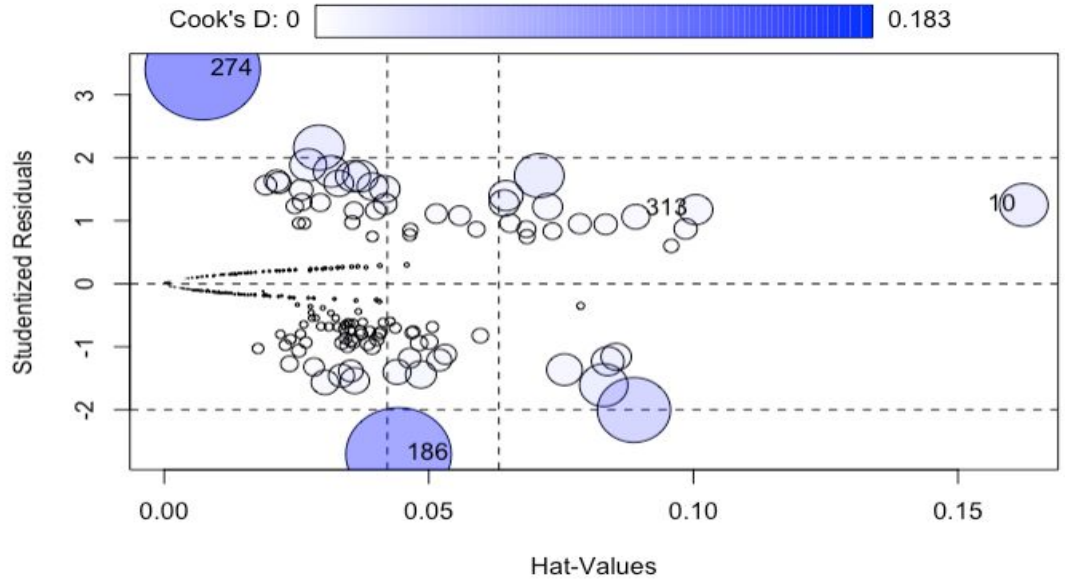


Figure 7: Residual Plot

Cross Validation for Log Reg Model

- Performed 10-fold cross validation
  - 70% of the data was used for training and 30% was used for testing
  - Performance Metrics:
    - Accuracy: 86.87%
      - 95% CI for Accuracy: (78.95%, 92.82%)
- After resampling the data 30 times:
  - Sensitivity: 0.9138
  - Specificity: 0.8775

		Actual		
		Below median	Above median	Column Totals
Predictions	Below median	44	6	50
	Above median	7	42	49
	Row Totals:	51	48	99

Figure 8: Confusion Matrix for Logistic Regression Model

ROC Curve & AUC

- The area under the curve is 0.9689804

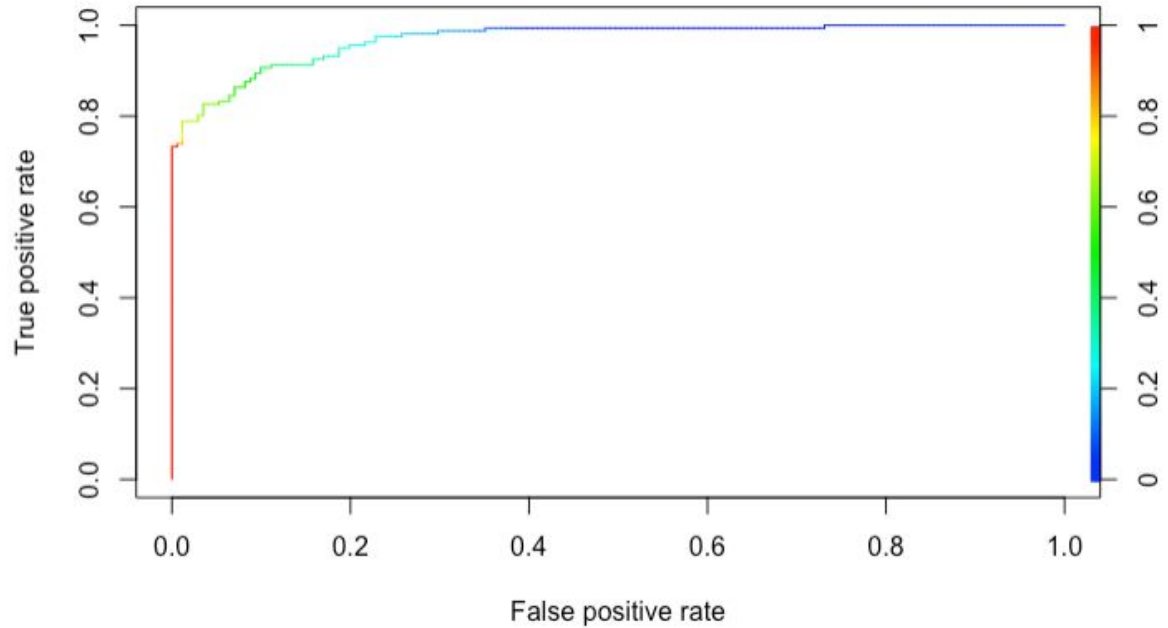


Figure 9: ROC Curve for Logistic Regression Model

Conclusion

- Our logistic regression model was a good fit for the data based on the Pearson's Chi-Square test, marginal model plots, and residual plots.
- After cross validation, our model showed a high accuracy of **86.87%**. The ROC curve confirmed the strength of our model due to the high AUC of **0.969**.
- Chinstrap penguins are less likely to be in the “above median” body mass group in comparison to Adelie penguins.
    - On the other hand, Gentoo penguins are much more likely to be in the “above median” body mass group than Adelie penguins.
- Males are 15.5 times more likely to be in the “above median” body mass group than females.
- Therefore, logistic regression is an effective method in determining whether body mass of Antarctic penguins can be predicted from species, bill length, bill depth, flipper length, and sex.

Shortcomings

- Consider exploring more interactions between variables
- Find a larger dataset
  - The dataset was limited to 3 penguin species inhabiting 3 islands in Antarctica, so expanding the scope of the data in terms of geography covered, the number of individual penguins covered, and the number of species of penguins covered would be ideal for further study.
  - More variables, categorical or numerical, could help us strengthen our model.
    - Example: age, eating habits, breeding
- When making body mass into a categorical variable, we decided to make the cut at the median. However, in the real world, the principal investigator of the study would decide where to make the most appropriate breaks.
  - If suggested, we can split our body mass variable into quartiles and consider exploring multinomial regression.

References

1. *Penguins*. BirdLife International. (2021, November 1). https://www.birdlife.org/birds/penguins/#:~:text=Penguins%20are%20sadly%20one%20of,as%20either%20Vulnerable%20or%20Endangered.
2. Carry-over body mass effect from winter to breeding in a resident ... (n.d.). https://royalsocietypublishing.org/doi/10.1098/rsos.140390
3. Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/. doi: 10.5281/zenodo.3960218.

Thank You!