Penguin Data Analysis: Predicting Body Mass and Species Classification

Xinyue Ha, MA Statistics
SDS 4392 Final Project, Instructor: Sami Cheong

Washington University in St.Louis

Abstract

This project focuses on penguin data analysis to explore body mass prediction problem and species classification problem. In this project, I use the dataset in the R package called `palmerpenguins`, which was collected near the Palmer Station in Antarctica. Then I applied Linear Regression Mixed Models to study body mass, and I utilized Multinomial and Binary Generalized Linear Models for species classification problem by using cleaned data with 333 rows and 8 variables. To sum up, this project not only revealed useful patterns in penguin traits but also enhance my ability to apply advanced linear modelling skills in ecological area.

1 Introduction

Penguins are one of the most interesting animals on the earth and also one of the important roles in the ecosystem of Antarctica. The data from the `palmerpenguins` R package were collected and made available by Dr. Kristen Gorman and the Palmer Station Long Term Ecological Research (LTER) Program. It describe the size measurements, clutch observations, and blood isotope ratios for adult foraging Adélie, Chinstrap, and Gentoo penguins observed on islands in the Palmer Archipelago near Palmer Station, Antarctica.

This dataset catched my eyes immediately for two main reasons. Firstly, I am a penguin lover and I still remember my first English novel is called Mr. Popper's Penguins, so finding the chance to work with real penguin data was undoubted exciting for me. Another important reason is that it is a clean, well-structured dataset, including both numerical variables (like flipper length and body mass) and categorical factors (species, island, and year) with relatively fewer layers. This combination makes it a strong candidate for applying multiple linear regression, generalized linear models, and mixed-effects modeling.

In addition to practicing statistical methods from class, I also hope to gain a better understanding of how real-world ecological data can be used to identify patterns. At the same time, I hope I can make little contributions to support ecological studies on penguins.

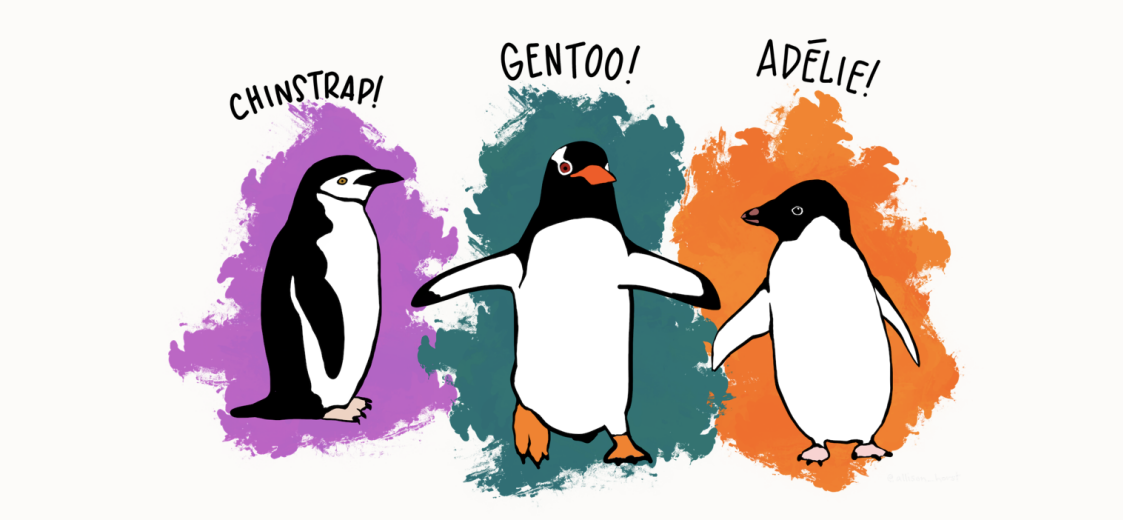

Figure 1: Palmer Penguins

2 Problem Statement

A. Predicting Body Mass

One core question I want to explore is how to predict a penguin's body mass based on its physical traits. Body mass is not only a basic biological measurement but also reflects a penguin's overall health, how well it can find food, and how prepared it is for breeding, all of which are important in ecological research. I planed to begin with a simple linear model to see if it could capture enough variation in mass using features and whether including grouping variables like year or island through a LLM would improve performance.

B. Species Classification

The second question is species classification. Accurate species identification helps scientist identify penguin species automatically, especially when working with large datasets or in places where it's hard to observe animals directly. This task is more challenging because it involves multiclass outcomes and some physical overlap between species. I planned to use a multinomial generalized linear model first, and then reframing the problem as a binary classification problem: Gentoo versus non-Gentoo to see if it can make the model more understandable.

3 About the Data

3.1 Variable Overview

The data used in this project come from the `palmerpenguins` R package, which offers a simplified and clean version of palmer penguins analysis. I chose to work with the simplified version (`penguins`), which includes 344 rows and 8 core variables, because these variables capture the essential biological and contextual information relevant to my analysis. Key variables include numeric measurements: *bill length*, *bill depth*,

flipper length, and *body mass*, as well as categorical variables: *species*, *sex*, *island*, and *year*. Together, they provide both the predictors and grouping factors necessary for building linear and classification models.

In contrast, the raw dataset (`penguins_raw`) contains 17 variables, including identifiers that are not directly related to my modeling goals. Such as scientific species names, sample numbers, and individual IDs and biochemical isotopes.

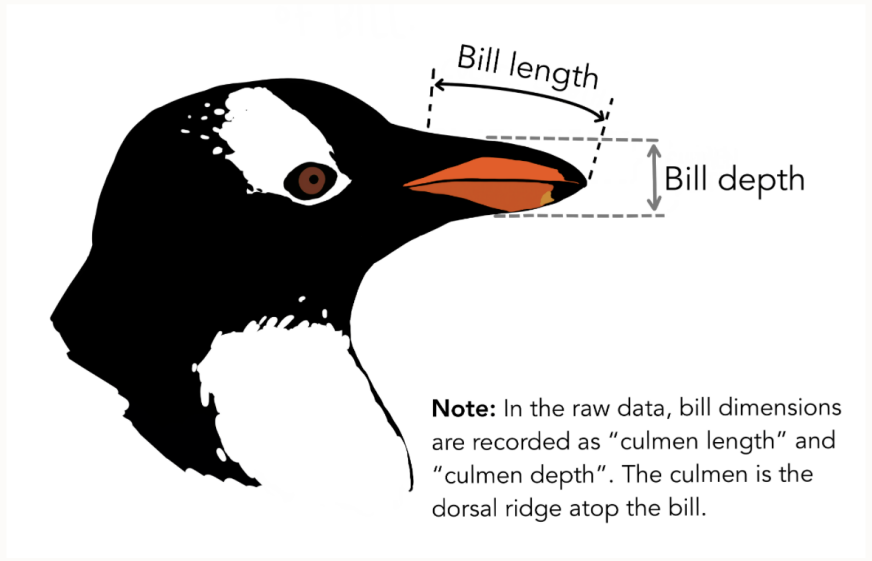

Figure 2: Difference in Bill length and Bill depth

3.2 Data Cleaning

I used `na.omit()` to remove 11 incomplete cases (3.2% of 344), mainly missing values in sex, flipper length, and bill depth. This left 333 complete observations for analysis. Then, I checked for outliers and pairwise correlations using boxplots and a correlation matrix, and found no major issues. Finally, I converted four categorical variables (species, island, year, and sex) into factors using `as.factor()` .

3.3 Graphical Summary

To observe the data structure clearly, I created several plots between variables. The first three continuous variables all show clear linear relationships with body mass. Among the categorical variables, sex and species show distinct differences in body mass, while island and year have smaller group effects. Since island and year reflect the group effects rather than biological traits, we can consider including them as random effects in a linear mixed model.

Then we can see the distribution of body features across species in Page 5 Figure 5. As shown in these four histograms, Gentoo penguins tend to have longer flipper length, smaller bill depth and heavier body mass compared to Adelie and Chinstrap. This suggests that Gentoo may be easier to distinguish using physical traits, while Adelie and Chinstrap have more overlap, which could make them harder to separate. For how to distinguish Gentoo, we will discuss later in 5.4.

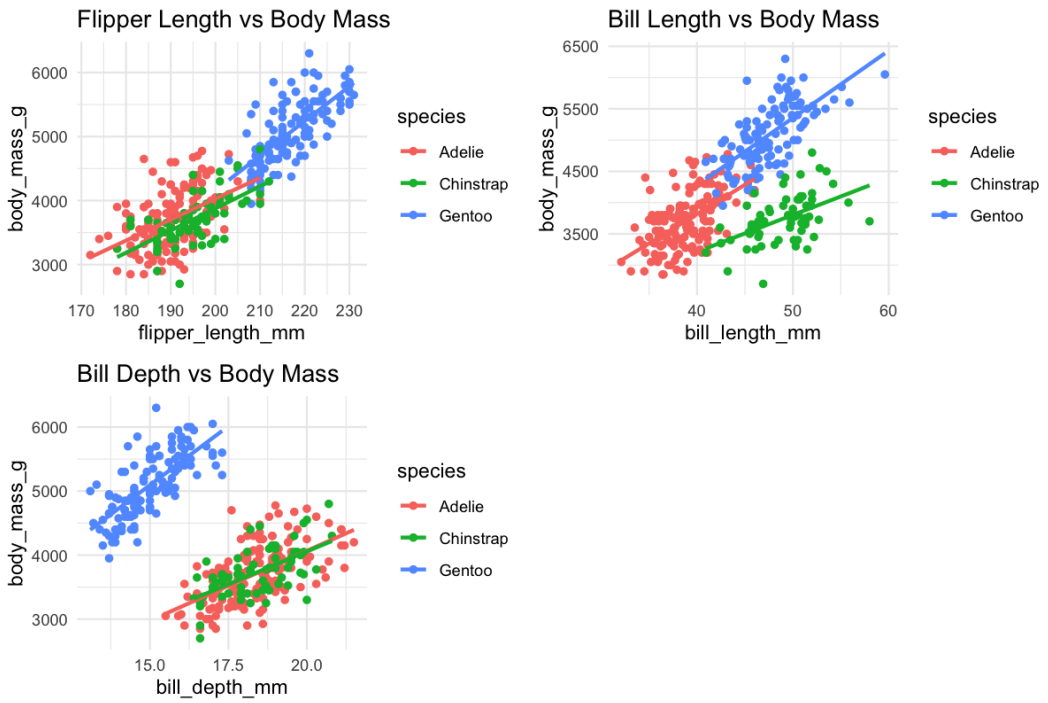

Figure 3: Scatterplots of body mass against flipper length, bill length, and bill depth, colored by species. These plots show clear linear trends and inter-species variation, especially for Gentoo penguins.

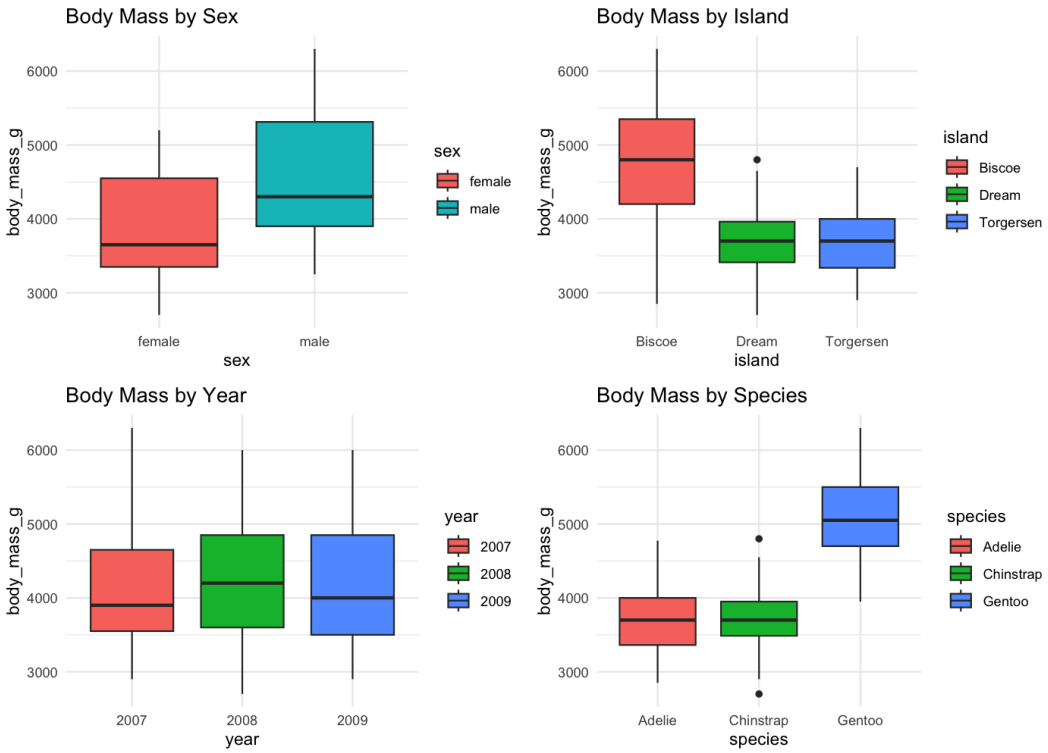

Figure 4: Boxplots of body mass by sex, island, year, and species. The species and sex variables show strong separation, while island and year contribute smaller group-level differences.

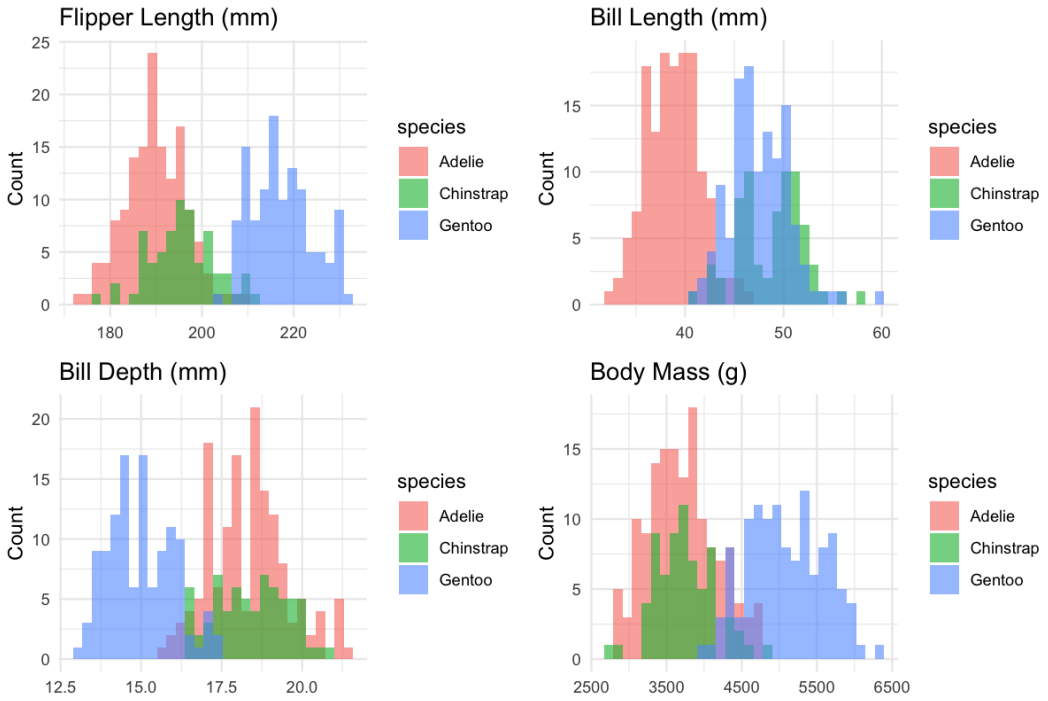

Figure 5: Histograms of continuous variables (bill length, bill depth, flipper length, body mass) by species. Gentoo penguins show clearly different distributions compared to Adelie and Chinstrap.

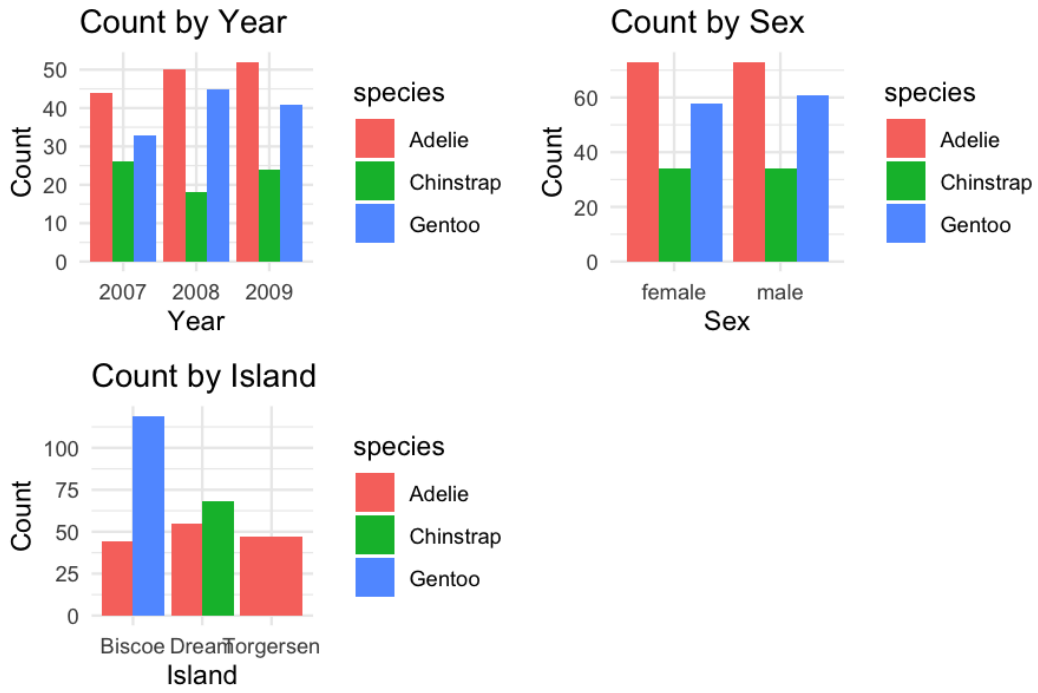

Figure 6: Bar plots of categorical variables (year, sex, island) by species. Distribution patterns support the potential for grouping effects in modeling.

4 Predicting Body Mass

4.1 Linear Regression Model

To find how penguin body measurements relate to body mass, I first fitted a multiple linear regression model with predictors including flipper length, bill length and depth, species, sex, island, and year. The model explained 87.3% of the variance in body mass. Species, sex, and flipper length were statistically significant predictors, suggesting these traits are biologically relevant, while residual plots showed no major deviations from normality.

However, island and year were not significant as fixed effects. Since they are group level factors rather than individual traits. Therefore, I considered modeling them as random effects in the next part, and trying to fitt a linear mixed model to see if it can help us better account for variability across groups.

```
model_lm <- lm(body_mass_g ~ species + bill_length_mm + bill_depth_mm +
               flipper_length_mm + island + sex + year,
               data = penguins_clean)
```

Predictor	Estimate	Std. Error	p-value
Species: Chinstrap	-281.0	89.1	0.0018**
Species: Gentoo	886.7	145.8	<0.001***
Bill Length (mm)	18.7	7.2	0.0097**
Bill Depth (mm)	60.3	20.1	0.0030**
Flipper Length (mm)	18.7	3.2	<0.001***
Sex (male)	379.6	48.2	<0.001***

Table 1: P-values for Linear Model

Residual diagnostic plots and the Shapiro-Wilk normality test indicated that residuals were approximately normally distributed ($W = 0.998$, $p = 0.9104$).

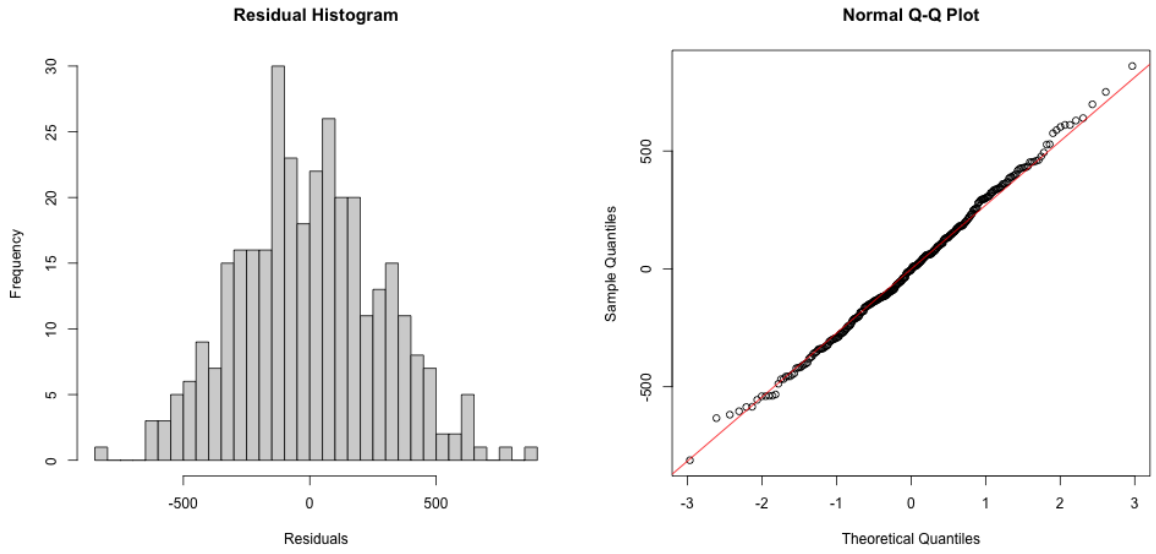

Figure 7: Residual diagnostic plots: Residual histogram (left) and Normal Q-Q plot (right). Both indicate residual normality.

4.2 Linear Mixed Model

After finding that `island` and `year` were not significant in the LM, I considered fitting a linear mixed model (LMM). I want to properly account for potential group-level variability introduced by `island` and `year` as random effects, rather than fixed predictors.

The fitted linear mixed model included the same fixed effects as before (`species`, `bill length`, `bill depth`, `flipper length`, and `sex`),with random intercepts for `island` and `year`:

```
model_lmm <- lmer(body_mass_g ~ species + bill_length_mm + bill_depth_mm +
                  flipper_length_mm + sex + (1 | island) + (1 | year),
                  data = penguins_clean)
```

Interestingly, the model returned a singular fit warning. According to the model output, the estimated random effect variance for `island` was exactly 0, while the variance for `year` was small but non-zero (784.8, Std.Dev = 28.01). This suggests that `island` contributes no measurable group-level variation, while `year` contributes very little.

Grouping Factor	Variance	Std. Dev.
Island (Intercept)	0.0	0.00
Year (Intercept)	784.8	28.01
Residual	82098.0	286.53

Table 2: Summary of Random Effects from linear mixed model.

The fixed effects estimates such as `species`, `sex`, `flipper length`, `bill length`, and `bill depth` remains highly significant.

Given the minimal variance from the random effects, especially the negligible contribution from `island`, I conducted the following three methods to evaluate if they were necessary:

- **Likelihood ratio tests**: Removing `island` or `year` did not significantly affect model fit ($p = 1.0$ and $p = 0.8675$).
- **Kenward-Roger ANOVA**: All fixed effects remained significant and with particularly strong contributions from `sex`, `flipper length`, and `species`.
- **Bootstrap test**: Provided similar results, confirming that a simpler model without random effects is adequate.

Thus, although the LMM provided valuable insight into the structure of the data, the simpler fixed effect linear model was actually sufficient for predicting penguin body mass reliably.

5 Species Classification

5.1 Model Overview: Multinomial Logistic Regression

Multinomial logistic regression helps generalize the binary logistic model to cases with response variable has more than two categories. Here, we can try this method for the response `species`, which includes three classes: Adelie, Chinstrap, and Gentoo.

The model estimates the log-odds of each class comparing with a baseline category (here is Adelie by default). For example, the probability of a penguin being Chinstrap or Gentoo can be modeled as:

$$\log\left(\frac{P(Y = \text{Chinstrap})}{P(Y = \text{Adelie})}\right) = \beta_0^{(1)} + \beta_1^{(1)} x_1 + \cdots + \beta_p^{(1)} x_p$$
$$\log\left(\frac{P(Y = \text{Gentoo})}{P(Y = \text{Adelie})}\right) = \beta_0^{(2)} + \beta_1^{(2)} x_1 + \cdots + \beta_p^{(2)} x_p$$

This is similar in structure to the binary logistic regression model used in GLMs with a binomial family:

$$\log\left(\frac{P(Y = 1)}{P(Y = 0)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

In our case, this method is more convenient for us to compare three species together.

5.2 Multinomial Logistic Regression Models

I first fitted a full multinomial logistic regression model with the response variable `species` and all the predictors:

```
model_multinom <- multinom(species ~ bill_length_mm + bill_depth_mm +
                           flipper_length_mm + body_mass_g +
                           island + sex + year,
                           data = penguins_clean)
```

To better assess significance, I computed z-scores and p-values by standard normal approximation:

```
z_vals <- summary(model_multinom)$coefficients /
          summary(model_multinom)$standard.errors
p_vals <- 2 * (1 - pnorm(abs(z_vals)))
round(p_vals, 4)
```

Predictor	Chinstrap (p)	Gentoo (p)
Bill Length (mm)	0.9098	0.5892
Bill Depth (mm)	0.9066	0.7174
Flipper Length (mm)	0.9461	0.9726
Body Mass (g)	0.9997	0.9818
Island: Dream	0.1804	**0.0000**
Sex: Male	0.3574	**0.0000**
Year 2008	0.0000	**0.0000**
Year 2009	0.6676	**0.0000**

Table 3: Full model: p-values for predictors across species.

From these results, variables such as `island`, `sex`, and `year` showed highly significant associations with species classification, especially for distinguishing Gentoo. However, this performance seems stem from **data collection bias**. For example, Gentoo penguins were almost exclusively observed on Biscoe island. It was a pity that these predictors mainly capture sampling related structure but not biological differences.

5.3 Body-only Model

To emphasize morphological distinctions and improve generalizability, I re-fit the model using only physical traits:

```
model_bodyonly <- multinom(species ~ bill_length_mm + bill_depth_mm +
                           flipper_length_mm + body_mass_g,
                           data = penguins_clean)
```

Predictor	Chinstrap (p)	Gentoo (p)
Bill Length (mm)	0.4227	**0.0000**
Bill Depth (mm)	**0.0969**	**0.0000**
Flipper Length (mm)	0.8664	**0.0000**
Body Mass (g)	0.7034	0.9960

Table 4: Body-only model: clearer signal for Gentoo species.

Compared to the full model, the body-only version revealed stronger biological signals—particularly for Gentoo penguins, which appear morphologically distinct. However, Chinstrap and Adelie remain harder to separate using body traits alone, consistent with our earlier exploratory plots. Therefore, this result motivates a potential reframing of the task as binary classification (Gentoo vs. non-Gentoo), which we explore next.

5.4 Binary Classification: Gentoo vs. Non-Gentoo

To simplify the species classification and reflect the structure observed in the data, we redefined the outcome as a binary variable to distinguish **Gentoo** penguins from the others. We created a new response variable using:

```
penguins_clean$gentoo_binary <- ifelse(penguins_clean$species == "Gentoo", 1, 0)
```

We then fitted two binomial logistic regression models using the `glm()` function with `family = binomial`:

```
model_bino_full <- glm(gentoo_binary ~ bill_length_mm + bill_depth_mm +
                       flipper_length_mm + body_mass_g + island + sex + year,
                       data = penguins_clean, family = binomial)

model_bino_body <- glm(gentoo_binary ~ bill_length_mm + bill_depth_mm +
                       flipper_length_mm + body_mass_g,
                       data = penguins_clean, family = binomial)
```

However, both models triggered convergence warnings:

```
Warning: glm.fit: algorithm did not converge
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

This suggests that the logistic regression encountered perfect separation. Below is the output from the full model:

Predictor	Estimate	Std. Error	p-value
(Intercept)	-137.8	1330000	1.000
bill_length_mm	0.6349	10400	1.000
bill_depth_mm	-8.85	37430	1.000
flipper_length_mm	0.9363	4836	1.000
body_mass_g	0.0161	105.5	1.000

Table 5: GLM summary output from the full model. All coefficients show inflated standard errors.

We observed similar results in the body-only model, where all p-values were also near 1. This strongly suggests that the GLM was unable to properly estimate model parameters due to near-perfect separation between groups.

To better understand this separation, we visualized the distribution of physical measurements by Gentoo status:

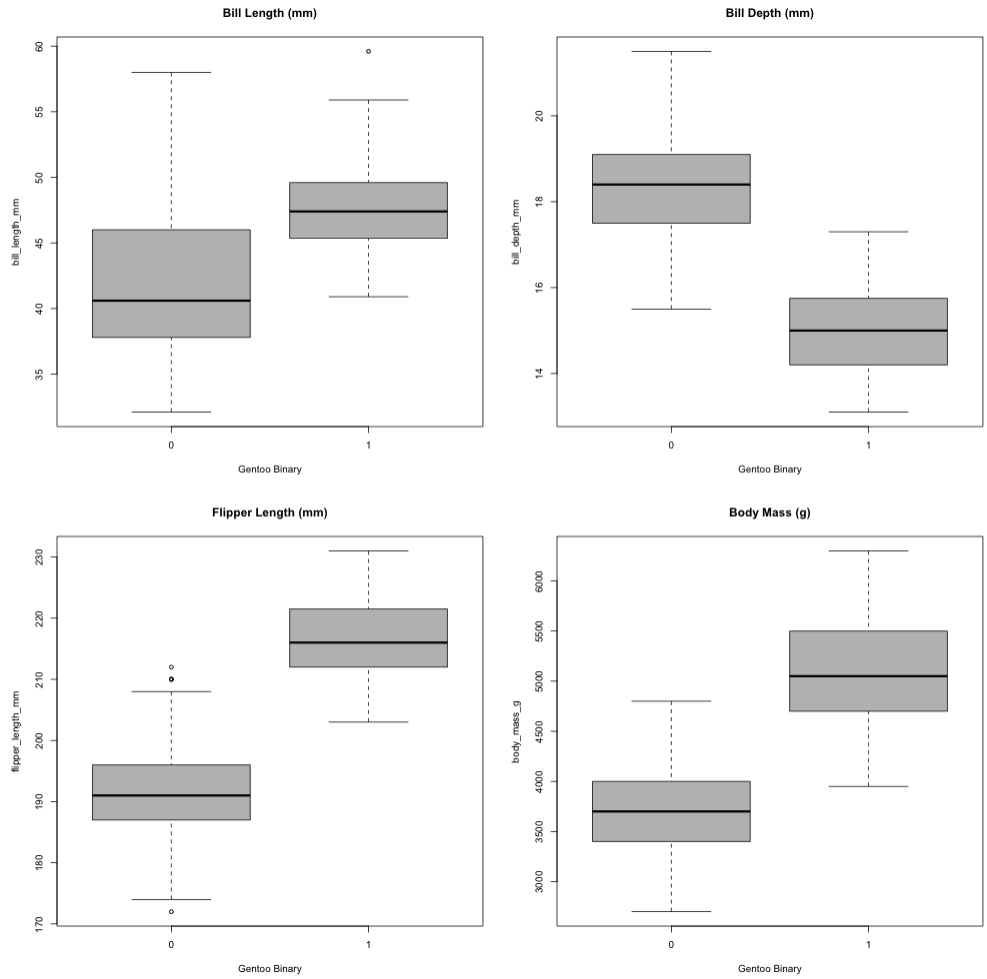

Figure 8: Boxplots of bill length, bill depth, flipper length, and body mass across Gentoo and non-Gentoo penguins. Gentoo penguins show strong morphological separation.

From Figure 8, it is clear that Gentoo penguins differ substantially in all four body measurements, especially in flipper length and body mass. These distinct group differences likely caused the perfect separation and convergence failure.

To address this, we refitted the model using Firth's penalized logistic regression via the `logistf()` function. This method is known to mitigate separation issues and provide bias-reduced estimates. The code and model output are shown below:

```
library(logistf)
model_firth <- logistf(gentoo_binary ~ bill_length_mm + bill_depth_mm +
                       flipper_length_mm + body_mass_g,
                       data = penguins_clean,
                       control = logistf.control(maxit = 1000))
summary(model_firth)
```

Predictor	Estimate	Std. Error	p-value
bill_length_mm	0.0291	0.1588	0.0125*
bill_depth_mm	-1.495	0.4210	<0.001***
flipper_length_mm	0.1392	0.1086	0.2053
body_mass_g	0.0020	0.0016	0.4376

Table 6: Firth logistic regression results using `logistf`.

Only `bill_length_mm` and `bill_depth_mm` were statistically significant in the Firth model, confirming that bill morphology is the most discriminative trait for identifying Gentoo penguins. Other features like flipper length and body mass showed consistent directional effects, although their p-values were not statistically significant under penalized estimation.

While the regular GLM encountered convergence issues due to near-perfect separation, it still revealed the strong discriminatory power of body metrics in distinguishing Gentoo penguins. The penalized logistic regression served as a robust complement, stabilizing parameter estimation and confirming the importance of key predictors while preserving interpretability.

6 Conclusion

In this project, I explored two main ecological questions. Using the penguin dataset collected from Palmer Station to predict penguin body mass and classify penguin species.

For body mass prediction, a linear regression model with only fixed effects proved sufficient. While the island and year variables considered as random effects did not make meaningful contributes. Nevertheless, I have another interesting consideration may be expanded in future analysis is that: If we can treat species as a random effect and design random slope models, such as `(1 + bill_length_mm | species)`. This might provide some insights into relationships between physical traits and body mass. However, such models may also face estimation challenges because of limited samples.

Regarding species classification problem, we revealed clear morphological distinctions, especially for identifying Gentoo penguins. Although logistic regression (GLM) encountered convergence difficulties due to perfect separation, the substituted penalized methods Firth's logistic regression made sense. From my perspective, the future directions for this kind of classification task could include some non-parametric methods or advanced machine learning algorithms like decision trees, random forests, or even trying feature engineering learning skills. Moreover, updating the dataset from recent years to get more additional observations to improve our models or verify the current findings is also of great significance.

Last but not least, I want to express my sincere appreciation for the efforts from ecological scientists and workers who collected this valuable dataset. Penguins will always hold a special place in my heart, and I hope this analysis can not only deepened my understanding in statictical learning, but also provided practical tools to make little contributions for the conservation of these remarkable lives in Antarctica, in such a small way.

7 References

The data were collected and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network.

The artworks were illustrated by @allison_horst](https://github.com/allisonhorst@allison_horst).

Gorman KB, Williams TD, Fraser WR (2014). Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (genus *Pygoscelis*). *PLoS ONE*, 9(3): e90081. https://doi.org/10.1371/journal.pone.0090081]

Palmer Station Antarctica LTER and K. Gorman (2020). Structural size measurements and isotopic signatures of foraging among adult male and female Ad'elie penguins (*Pygoscelis adeliae*) nesting along the Palmer Archipelago near Palmer Station, 2007–2009 ver 5. Environmental Data Initiative. https://doi.org/10.6073/pasta/98b16d7d563f265cb52372c8ca99e60f]

Figure 9: Logo of Palmer Penguins