Using Linear Regression to Predict Body Mass in Penguins from Antarctica

Group 6: Emiily Schwartz, Matthew Weiss, and Andrew vanderWilden

Contents

1 Abstract

Ecological studies are crucial to understanding animal populations. This report uses data from Dr. Kristen Gorman and the Palmer Long-Term Ecological Research Observatory. The data describe three species of penguins observed on three islands in the Palmer Archipelago, Antarctica. We use multiple linear regression to estimate penguin body mass. We find sex, flipper length, and bill depth to be significant predictors in our model.

2 Introduction

2.1 Orientation Material

This report analyzes various traits of penguins to try to predict their body mass using linear regression. The data analyzed was collected and made available by Dr. Kristen Gorman and the Palmer Long-Term Ecological Research Observatory. The data describe three different species of penguins observed from three islands in the Palmer Archipelago, Antarctica.

The data are cross-sectional and describe various physical traits of the penguins.

2.2 Key Aspects

In this report, we fit a least-squares linear regression model to predict the body mass of penguins. We find the variables flipper length, sex, and bill depth to be significant predictors in our model.

2.3 Plan for the Rest of the Report

The outline for the remainder of the report is as follows. In section 3, we present the most important characteristics of the data and relationships between predictor variables and body mass. In section 4, the model selection process and following interpretation will be discussed. Concluding remarks can be found in section 5 with details to follow in the Appendix.

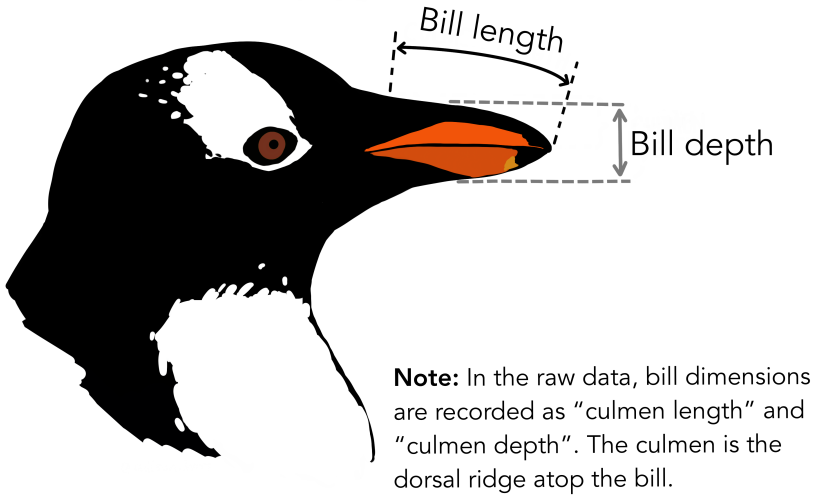
3 Data Characteristics

The data are cross-sectional and describe traits for 342 penguins observed at the Palmer Long-Term Ecological Research study area. The data contains information about the following 7 variables:

3.1 Variables

Item	Variable	Definition
1	species	Species of Penguin (Adelie, Chinstrap, Gentoo)
2	island	Island where Penguins was Observed (Biscoe, Dream, Torgersen)
3	bill_length_mm	Length of Penguin's Bill in Millimeters
4	bill_depth_mm	Depth of Penguin's Bill in Millimeters
5	flipper_length_mm	Length of Penguin's Flipper in Millimeters
6	body_mass_g	Penguin's Body Mass in Grams
7	year	Year Penguin was Observed and Marked

See the below picture for clarification on bill length and depth definitions:



The data contained 9 observations that were missing information on the sex of the penguin. In order to account for these missing values, we used recursive partitioning to impute the most likely sex of the penguin.

The variable of interest in this report is the `body_mass_g`, or the body mass of the penguins measured in grams.

We will now introduce and dive deeper into each of the variables in the dataset.

3.2 Summaries

The below table shows the five number summary for all numerical variables in the dataset:

Variable	Mean	Median	St. Dev	Max	Min
body_mass_g	4201.75	4050.00	801.95	6300.0	2700.0
bill_depth_mm	17.15	17.30	1.97	21.5	13.1
bill_length_mm	43.92	44.45	5.46	59.6	32.1
flipper_length_mm	200.92	197.00	14.06	231.0	172.0
year	2008.03	2008.00	0.82	2009.0	2007.0

We can also observe the distribution of observations for the three categorical variables in the data set:

```
     species          island       sex
Adelie   :151   Biscoe   :167   female:172
Chinstrap: 68   Dream    :124   male  :170
Gentoo   :123   Torgersen: 51
```

Upon further examination, we can see the penguins are not spread out in a geographically diverse manner:

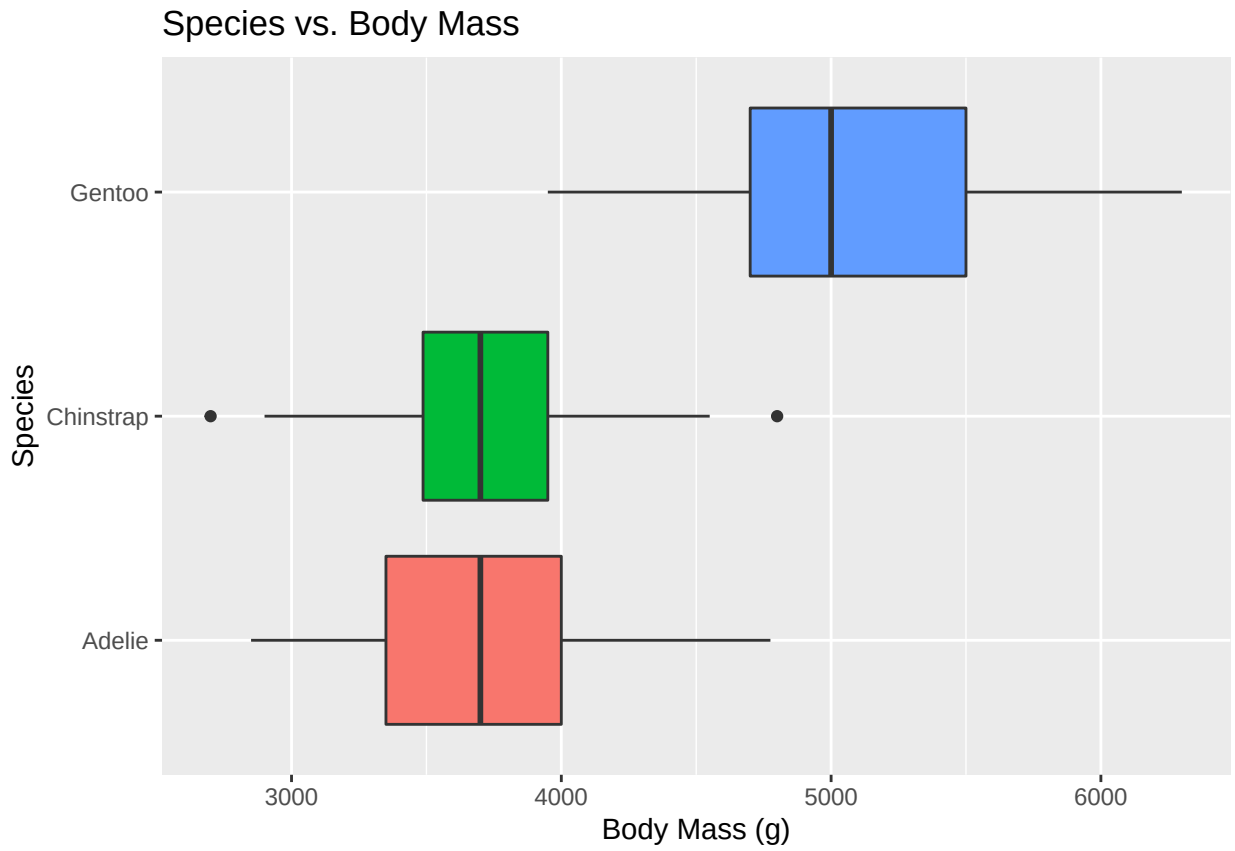
```
           Adelie Chinstrap Gentoo
Biscoe         44         0    123
Dream          56        68      0
Torgersen      51         0      0
```

We can observe from the above table that both Gentoo and Chinstrap penguins are observed on only one island while the Adelie penguins can be found on all three islands. This information is useful in that it shows much of the potential usefulness of the variable `island` is likely already captured in the variable `species` and likely won't need to be included in the model.

We will now examine each of the categorical variables more in-depth.

3.3 Species

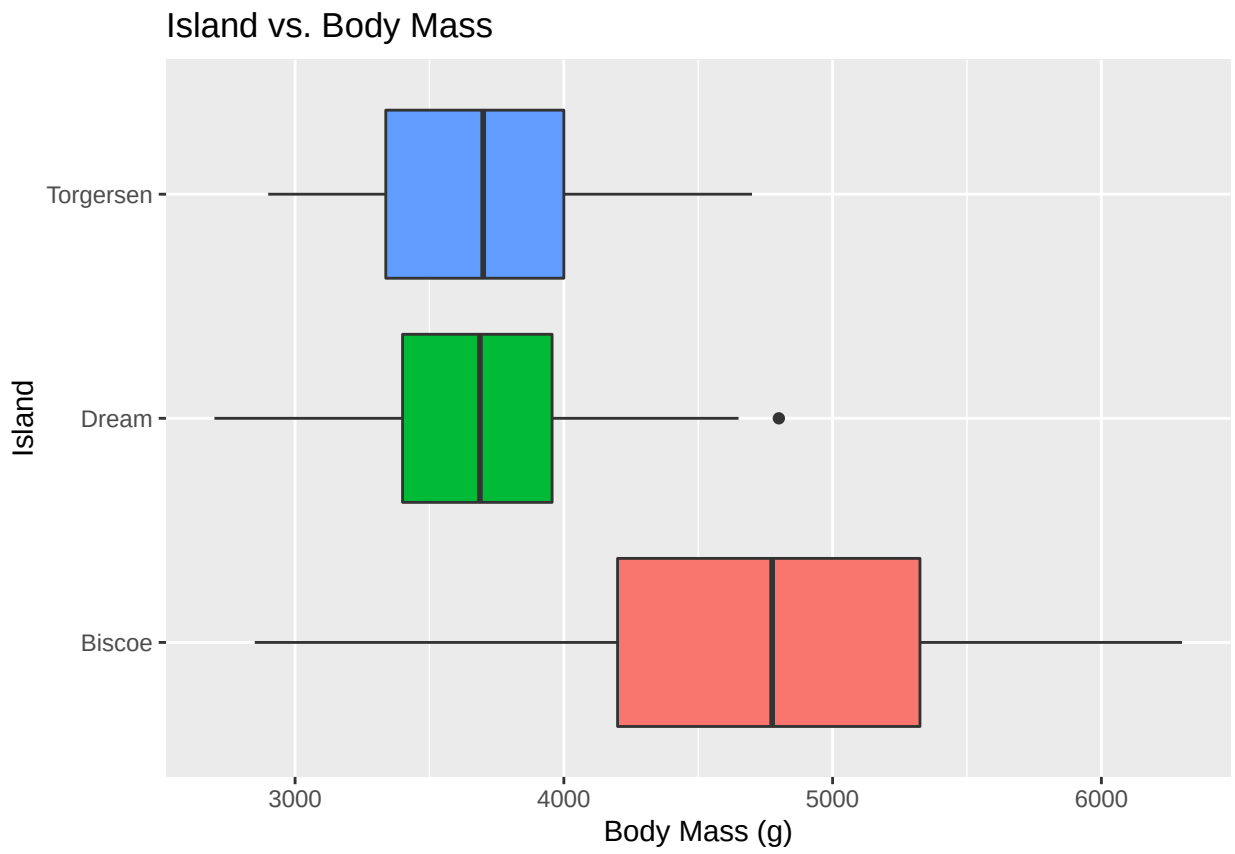
When examining a categorical variable, using boxplots can show differences between groups. As we can see from the below plot, Gentoo penguins have on average a larger body mass than the other two species of penguins



This suggests that `species` is likely a useful predictor of body mass, however it is possible this information will be captured by the inclusion of other variables.

3.4 Island

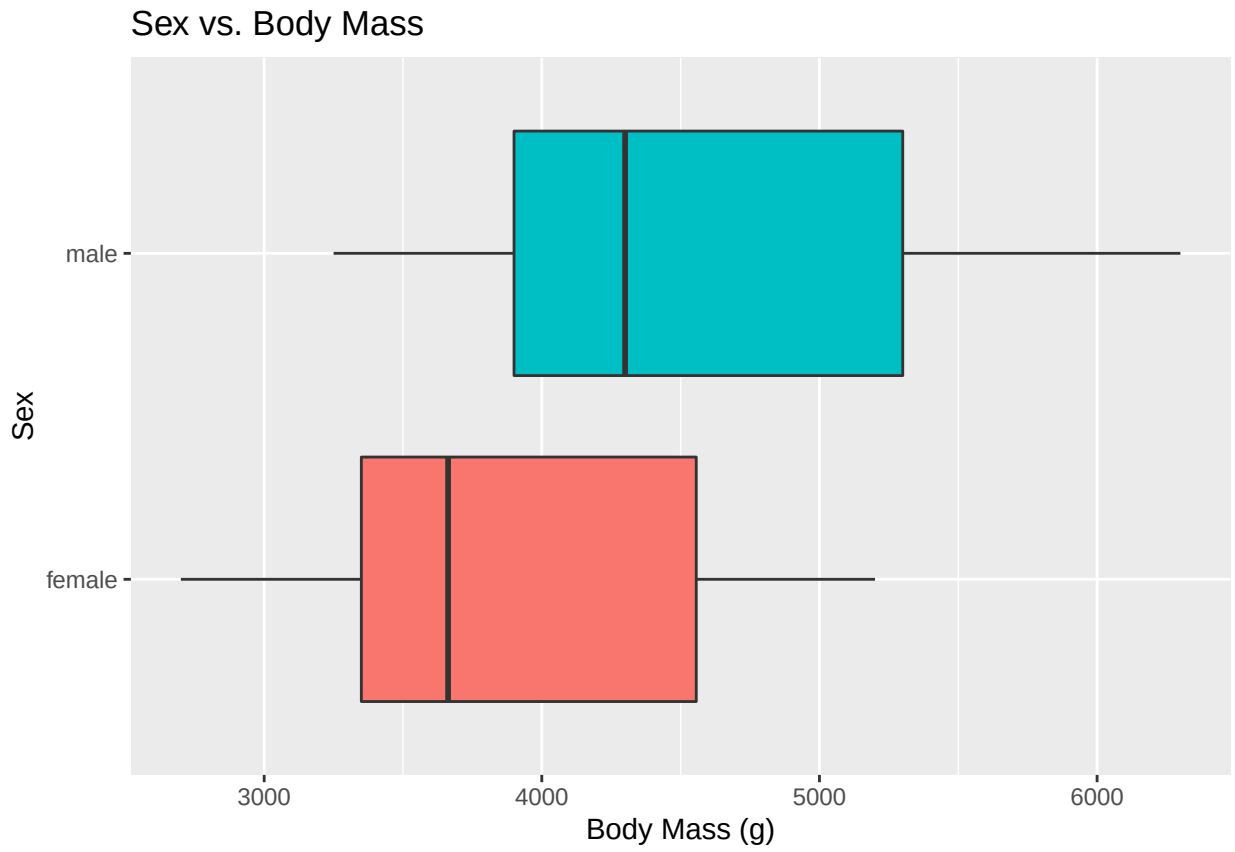
The below plot shows the average body mass of penguins broken down by island.



As noted previously, it appears there is a clear effect but this is likely because the penguins the species of penguins are not distributed in a geographically diverse manner.

3.5 Sex

The below plot shows the average body mass of penguins differentiated by sex:



We can clearly observe male penguins appear to have heavier body masses. This suggests the variable `sex` is useful in predicting body mass.

We will now examine the effect of the numerical variables.

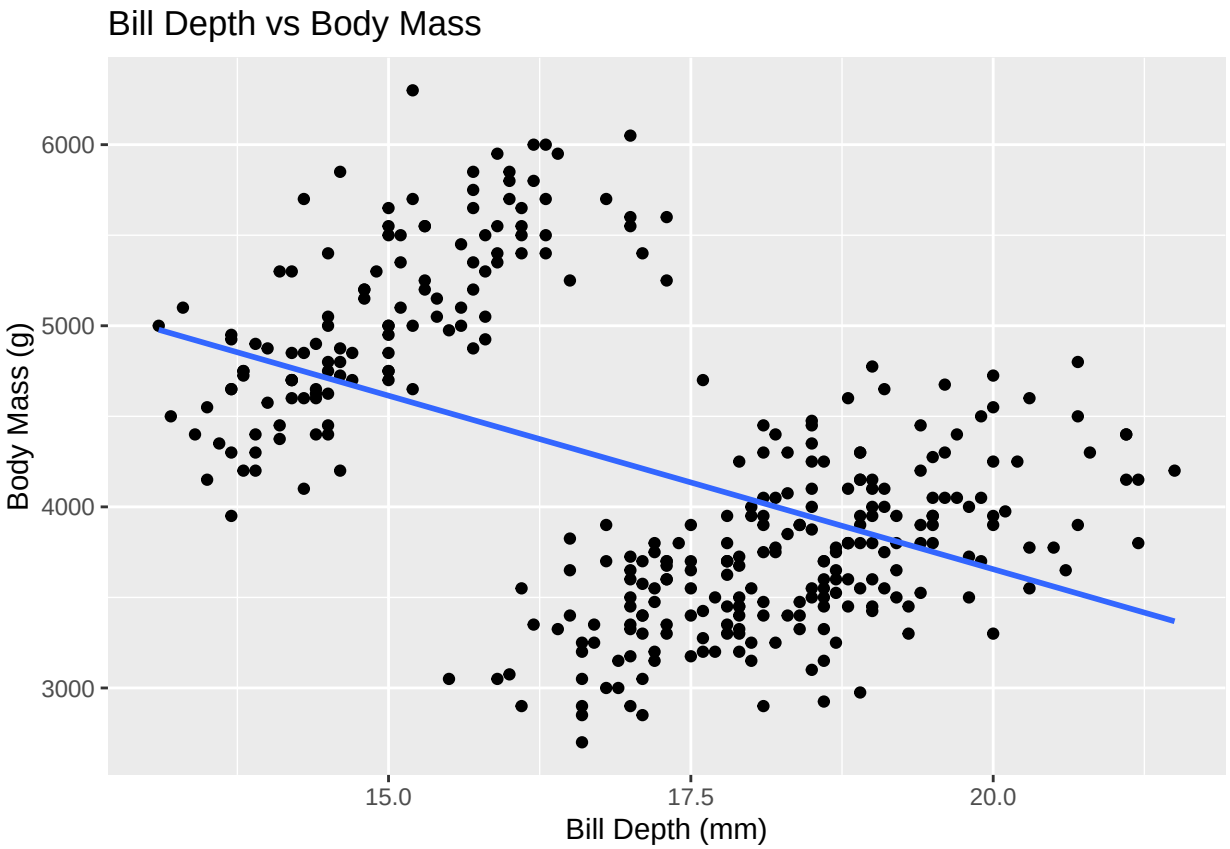
3.6 Bill Length

The below plot shows bill length plotted against body mass:

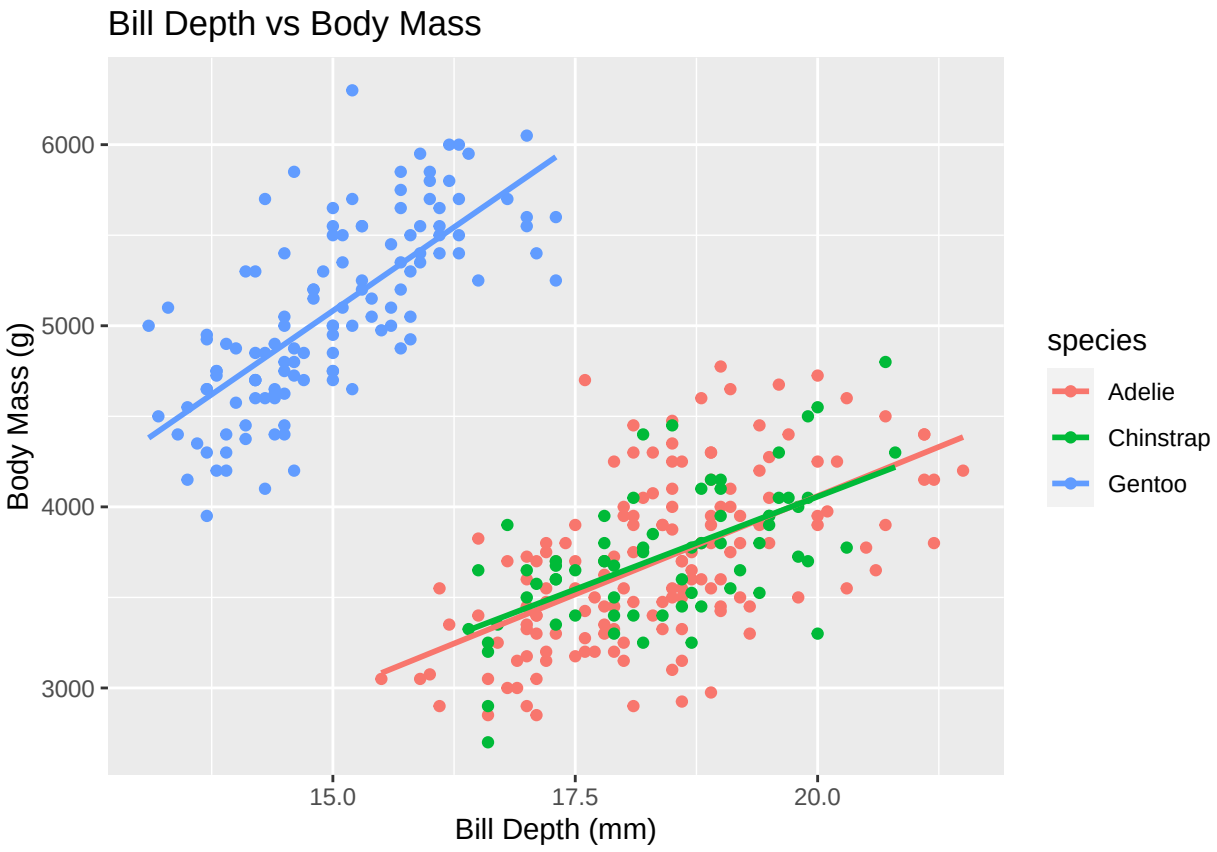
We can observe both from the data and the trend line there appears to be a clear positive linear relationship between the two variables. This suggests the variable `bill_length_mm` is a useful predictor of body mass.

3.7 Bill Depth

The below plot shows bill depth plotted against body mass:



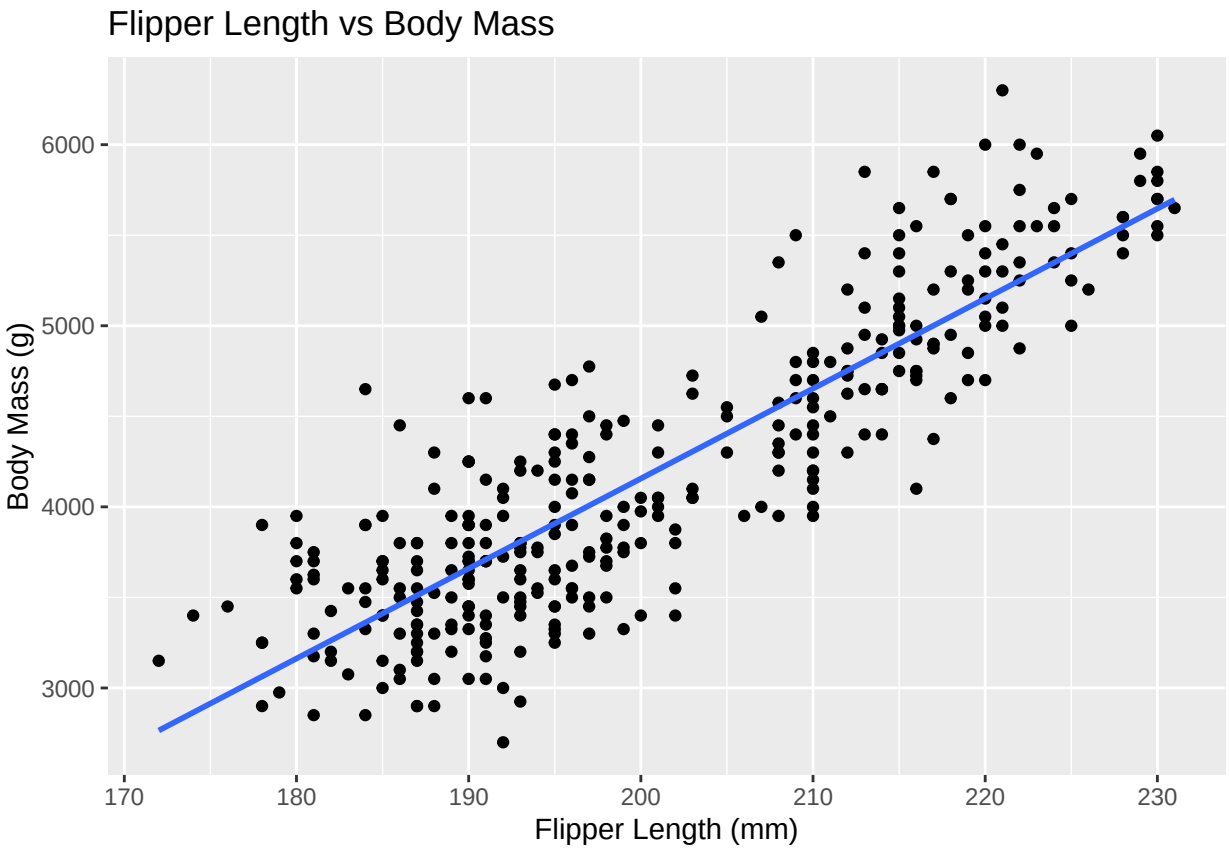
The trend line appears to show a negative linear relationship. This information could be useful however it could also tell a different story. We plotted the same graph but differentiated by species:



This graph tells a different story, showing a positive linear relationship between the two variables. Both pieces of information suggest there is a relationship between bill depth and body mass, suggesting the variable could be of use as a predictor.

3.8 Flipper Length

The below plot shows flipper length plotted against body mass:



There appears to be a clear positive linear relationship between flipper length and body mass, suggesting `flipper_length_mm` is likely a useful predictor of body mass

4 Model Selection and Interpretation

Based on the above data characteristics section, it has been established there are clear correlations and patterns between the the body mass of penguins and the various covariates.

In this section we summarize these relationships using regression modeling.

Our model was fit using 282 randomly selected observations with 60 observations withheld for model accuracy testing.

A number of models were considered. Many more complex models that used either polynomials or interaction terms suffered from multi-colinearity problems. This is likely due to some of the factors outlined in the data characteristics section which highlighted much of the information was redundantly captured across multiple variables. This led us to select a more simplistic model however we feel it fits the data well and offers valuable insights.

Based on our investigation of the data, we recommend a linear regression model to estimate the body mass of penguins in the Palmer Archipelago, Antarctica. The variables used to create the model are `flipper_length_mm`, `sex`, and `bill_depth_mm`. The variable `sex` contained 9 missing values. These values were imputed using recursive partitioning to assign the most likely sex of either male or female to the penguins.

4.1 Model

$$BodyMass_i = \beta_0 + \beta_1 FlipperLength_i + \beta_2 Sex_i + \beta_3 BillDepth_i + \epsilon_i$$

4.1.1 Coefficients

Term	Estimate	Std. Error	t Statistic	P value
(Intercept)	-2418.295	678.28	-3.565	0.0004276
flipper_length_mm	38.697	2.24	17.266	0.0000000
sexmale	540.979	56.23	9.620	0.0000000
bill_depth_mm	-84.139	17.04	-4.937	0.0000014

Both Flipper Length and Bill Depth are numerical covariates however sex is a categorical variable. The base level of the variable sex is taken to be female, while a male penguin is assigned a value of 1.

We can observe all four p-values of the coefficients appear to be significant at the $\alpha = 0.05$ level. In this case, the intercept term does not have a practical interpretation as it would describe a penguin without any flipper or bill. There are no penguins with a negative body weight.

$$\widehat{BodyMass} = -2418.295 + 38.697 FlipperLength + 540.979 Sex - 84.139 BillDepth$$

Now, let us interpret the model's coefficients from the table above. Looking at flipper length, as flipper length increases by one millimeter, the body mass will increase by approximately 38.697 grams. If male, one could predict the body mass of the penguin would increase by 540.979 grams. If female, this coefficient would equal 0 and therefore cancel out this 540.979 grams. In other words, if female, the body mass of the penguin would not increase based on this model. When looking at the bill depth of a penguin, if the bill depth increases by one millimeter, we could expect the body mass of the penguin to decrease by 84.139 grams

For example, suppose we wanted to predict the body mass of a penguin with the following characteristics:

Sex	Flipper Length (mm)	Bill Depth (mm)
male	204	16.1

We would use the model as such:

$$4662.21 = -2418.295 + 38.697(204) + 540.979(1) - 84.139(16.1)$$

We would interpret this to mean we predict a male penguin with a bill depth of 16.1 mm and a flipper length of 204 mm would have a body mass of 4662.21 grams (see appendix for prediction and confidence intervals).

4.1.2 Goodness of Fit

Statistic	Value
R-squared	0.8297
Adj. R-squared	0.8278
AIC	4094.5778
F Statistic	451.3871
P Value	2.2e-16
Degrees of Freedom	3, 278

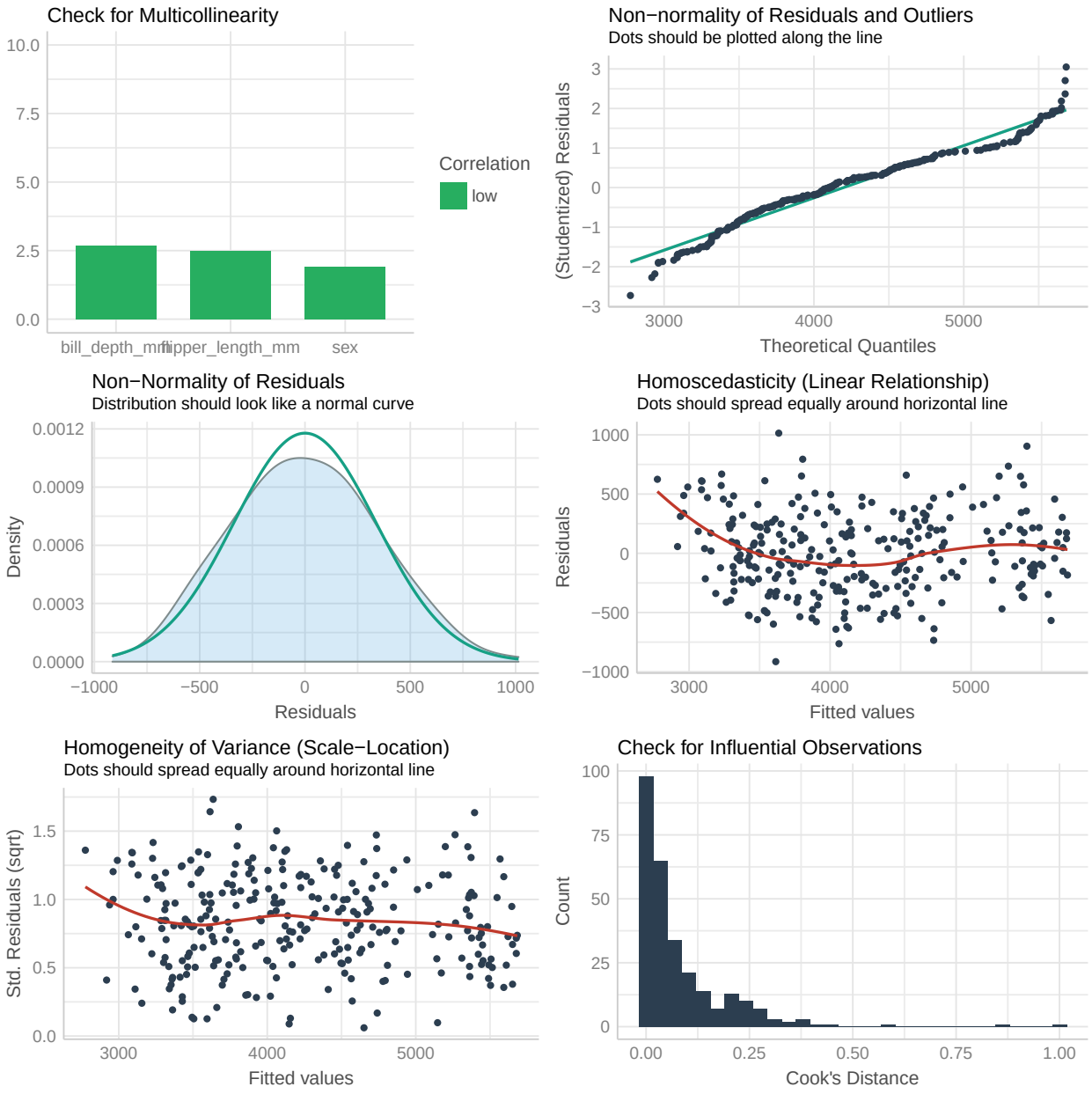
The R-squared value tells that 82.97% of the variance in the response, body mass, can be explained by the predictors, in this case the flipper length, bill depth, and sex of the penguin. We can also observe the model passes the global F-test as the p-value is less than the significance level of $\alpha = 0.05$, indicating the model is more useful in predicting body mass than an intercept-only model would be.

4.1.3 Performance

Build RMSE	Test RMSE
338.059	345.55

From the above table, we can observe the model performed nearly as well on data it had never seen before as it did on the data used to build the model. We would expect the Test-RMSE to be slightly worse than the build-RMSE but a large difference would suggest the model has been overfit to the data used to build it. The fact that the Test-RMSE is only slightly higher than the build RMSE suggests the model is not overfit.

4.1.4 Linear Regression Assumption Checks



Looking at the linear regression assumption checks above, let us key in on the normal probability plot and the residuals vs fits plot. For the normal probability plot, we can see that the data points showing the theoretical quantities' response on the residuals are all relatively plotted along the line, showing the normality of the plot. This also shows there is no large impact of outliers on the data and suggests a linear trend. Looking at the residual plot, which plot the fitted values versus the residuals, we can see that the data points are plotted about equally around the horizontal line, suggesting the error terms are normally distributed.

5 Summary and Concluding Remarks

In trying to predict the body mass of penguins in the Palmer Archipelago, Antarctica region, we found the variables `sex`, `flipper_length_mm`, and `bill_depth_mm` to be significant predictors in our linear regression model. This information is useful for the ecological study of penguins as well as monitoring long term trends in the penguin population as a whole. It is likely worth running similar analyses on penguins from other regions of the world to see if the findings are consistent or if they apply only to penguins from this region. It is also worth exploring if other methods of modeling would yield different results.

6 Appendix

95% prediction interval: 3986.8992244 $\leftrightarrow$ 5337.5209647

95% confidence interval: 4579.7000903 $\leftrightarrow$ 4744.7200988

6.1 Citations

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

Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/