Palmer Penguins

STAT 6730: Data Analysis Project

Frank Leyva Castro, Yousef Qaddura, Changrui Wang

Introduction

The `palmerpenguins` raw data were collected over 2007-2009 and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network. The cleaned-up dataset `penguins` contains complete data for size measurements of 333 adult foraging penguins of 3 different species (Adelie, Gentoo, Chinstrap), collected from 3 different Islands (Torgersen, Biscoe, Dream).

The dataframe `penguins` has `nrow(penguins)` complete cases and `ncol(penguins)` features. The following is the list of features present in the data along with a short description:

- Species `species`: the species of the penguin, a categorical variable with unique values Adelie, Gentoo, Chinstrap.
- Island `island`: which island the penguin is from, a categorical variable with unique values Torgersen, Biscoe, Dream.
- Bill Length & Depth `bill_length_mm, bill_depth_mm`: the length and depth measurements in mm of the penguin's bill, continuous variables. The length ranges from 32.1 to 59.6 with mean 43.9927928 and the depth ranges from 13.1 to 21.5 with mean 17.1648649.
- Flipper Length `flipper_length_mm`: the length of the penguin's flipper, a continuous variable ranging from 172 to 231.
- Sex `sex`: the sex of the penguin, a categorical variable with unique values male, female.
- Year `year`: the year in which the penguin's measurements were collected, a categorical variable with unique values .
- Body Mass `body_mass_g`: the penguin's mass measurement in grams, a continuous variable ranging from 2700 to 6300 with mean 4207.0570571.

Our questions of interest are three-fold:

1) How does body mass depend on the other features? More specifically, what are the best choices of linear regression models that explain body mass? Here best is in the sense of BIC, AIC and similar measures. This question is answered through best subsets regression.

2) Are the chosen models distinguished in terms of predictive power? We use 15-fold cross-validation to uncover the answer. We will find that all three have similar predictive power and hence choose to proceed with the simplest model as the model of choice.

3) What are the statistical properties of the coefficients for the model of choice? This question is answered using bootstrap.

Expolatory Data Analysis

Before embarking on our questions, we remark on our observations from Figure 1. Firth, we observe that changing sex from female to male remarkably causes an increase in the body mass axis ticks. Next, body mass is positively correlated with all size measurements. The Gentoo species (blue) is distinguished in having much higher body mass than other species. The other two species differ on their bill length. We also include an extra Figure 2 which shows much more statistical amongst all pairs of variables although we feel that it is enough to have commented on Figure 1 for the purposes of our analysis.

Best Subsets Regression Result

The results of best subsets regression are shown in Table 1 and Table 2. Aided with the observations in the previous section, we choose the simplest model to be the one with 4 variables in Table 1 so as to at least incorporate sex, Gentoo species and some size measurements. We choose the complex model with interactions to be the one with 5 variables in Table 1, although we enforce hierarchy. Note that Table 2 shows that all models with at least three variables have relatively comparable meaasure characteristics. We summarize the model equations:

- Simple (Model A)

$$\text{BodyMass} = \text{Male} + \text{Gentoo} + \text{Bill Depth} + \text{Flipper Length}$$

- Complex without Interactions (Model B)

$$\begin{aligned}\text{BodyMass} &= \text{Male} + \text{Gentoo} + \text{Chinstrap} \\ &\quad + \text{Bill Length} + \text{Bill Depth} + \text{Flipper Length}\end{aligned}$$

- Complex with Interactions (Model C)

$$\begin{aligned}\text{BodyMass} &= \text{Male} + \text{Gentoo} + \text{Chinstrap} \\ &\quad + \text{Bill Length} * \text{Bill Depth} + \text{Bill Length} * \text{Flipper Length}\end{aligned}$$

Anova Analysis

To analyze the models, begin by performing ANOVA F-tests. With Model A as null and B as alternative, the p-value of the F-test is given by 0.0087314. With Model B as null and C as alternative, the p-value of the F-test is given by 0.0188342. This shows that Model C best explains the variation in body mass compared to the other two models. This would be the model of choice if such characteristic is desired.

15-fold Cross-Validation

In this analysis, our model of choice will be based on its predictive power and simplicity. By performing 15-fold cross-validation (see Table 3), we observe that the three models have very similar average mean squared errors and comparable standard errors or mean squares across the folds. From this, we conclude that our model of choice is Model A, the simplest one as it is simplest with similar predictive power to the other two models.

Bootstrap Confidence Intervals

The bootstrap %95 confidence intervals for our model of choice coefficients are shown in Table 4. We observe that none of them contain the value zero and all have relatively acceptable widths. This entails that there is no strong evidence of bias toward model A in the data!

Conclusions

Overall, the *penguins* data-set is an excellent practice for newcomers into data analysis. The models and plots show clear species and sex differences and the remarkable dependence of a penguin's body mass on its bill and flipper characteristics. Model C highlighted the importance of interaction terms in explaining variance, but also gives a warning about the challenges of interpretation. All models considered have similar predictive power and the best model of choice recommended is the simplest!

Figure Appendix

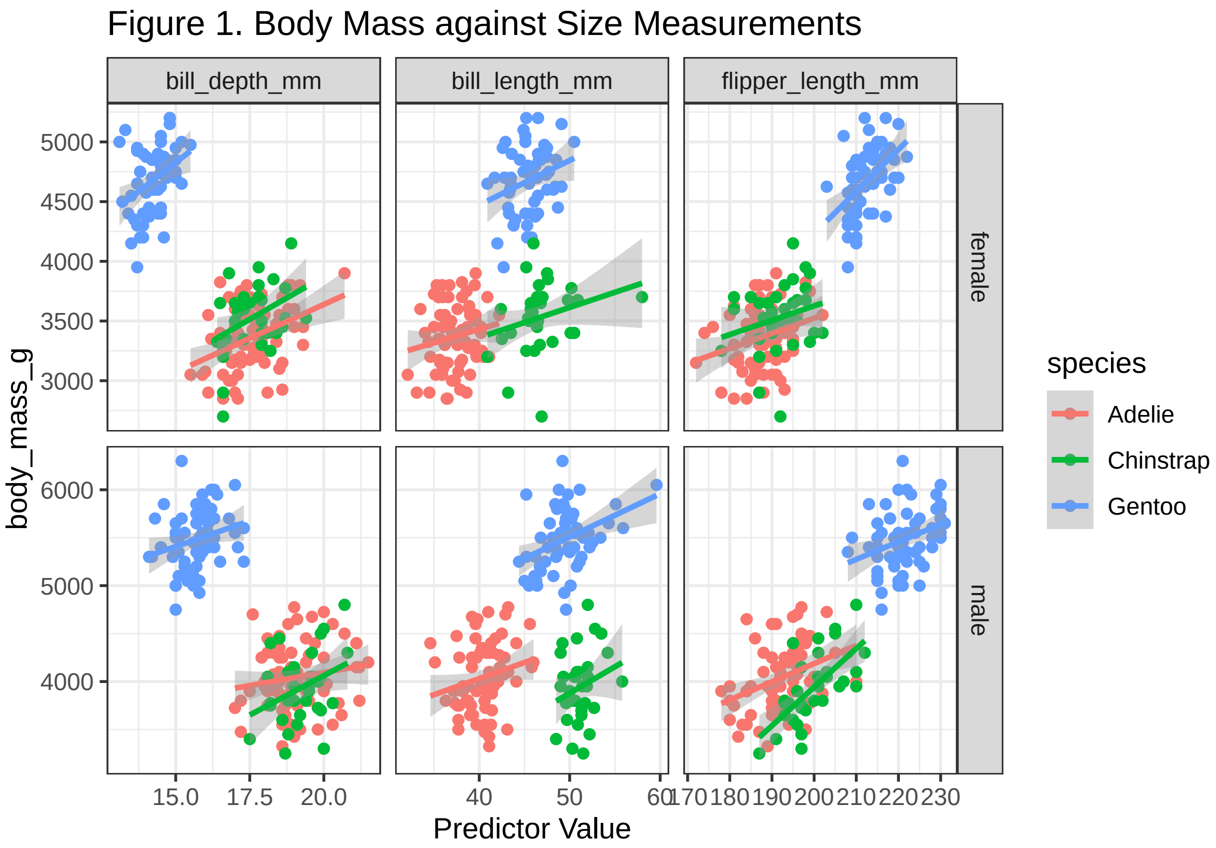

Figure 2. GGPairs Plot

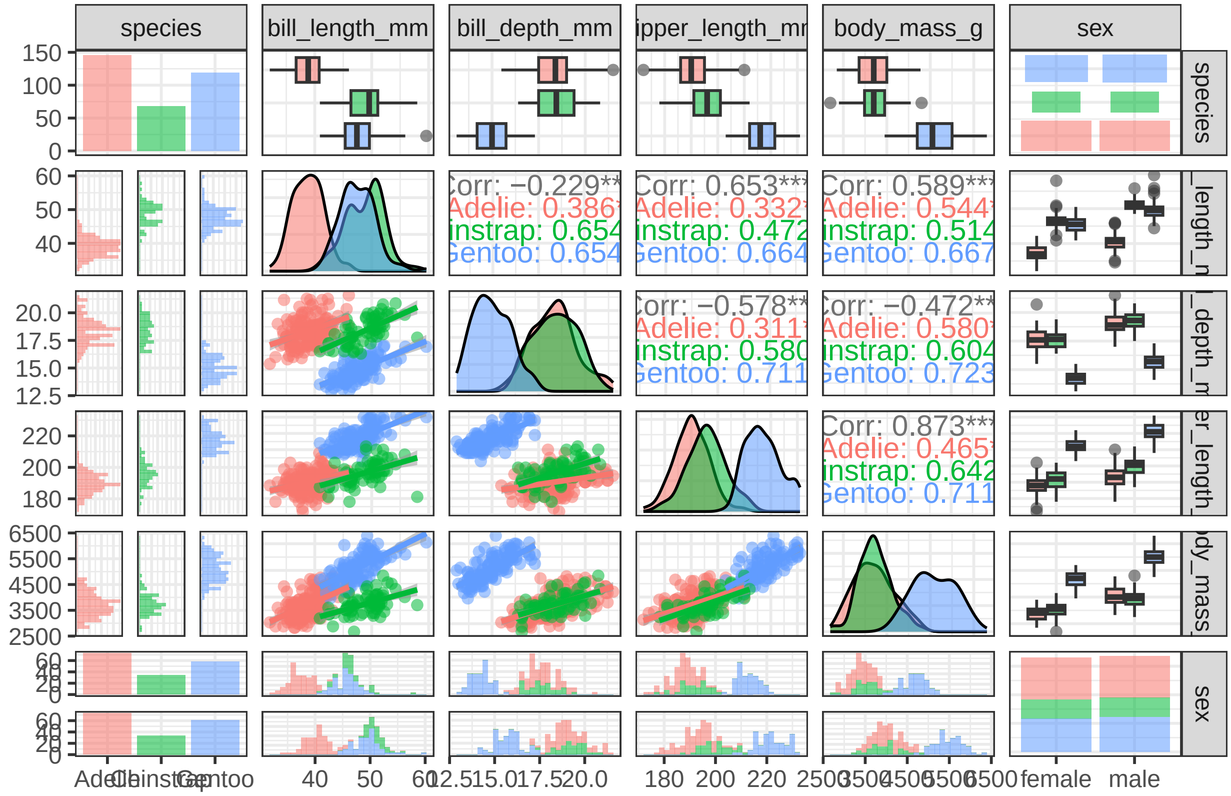

Table 1: Table 1. Best Subsets Regression Result

	male	Chinstrap	Gentoo	BillL	BillD	FlipperL	BD*BL	FL*BL	FL*BD	FL*BDBL
1 (1)						*				
2 (1)	*		*							
3 (1)	*		*						*	
4 (1)	*		*		*	*				
5 (1)	*		*		*		*	*		
6 (1)	*	*	*		*		*	*		
7 (1)	*	*	*	*	*		*		*	
8 (1)	*	*	*	*	*		*		*	*

Table 2: Table 2. Best Subsets Regression Measure Results

model.no	Mallow_C	BIC	R2	AR2	AIC
1	300.928653	-466.5291	0.7620922	0.7613734	1085.920
2	79.071488	-606.9337	0.8466372	0.8457078	1052.355
3	22.858018	-652.6522	0.8686230	0.8674250	1041.821
4	17.844381	-653.6268	0.8712718	0.8697020	1042.171
5	8.425038	-659.1663	0.8755846	0.8736823	1041.411
6	6.198781	-657.6579	0.8771808	0.8749203	1042.365
7	5.969181	-654.1407	0.8780229	0.8753957	1043.808
8	7.692176	-648.6183	0.8781275	0.8751183	1045.738

Table 3: Table 3. 15-fold cross-validation

Model	Mean Squared Error	Standard Error
A	85628.54	6899.618
B	84958.55	7240.633
C	85259.28	6470.569

Table 4: Table 4. Bootstrap Confidence Intervals for Model A Coefficients

	Estimated.Coefficients	Lower.Bound	Upper.Bound
maleTRUE	444.84014	347.5303	546.2287
gentooTRUE	1229.23800	1042.6639	1415.1893
bill_depth_mm	76.12165	38.0325	119.1671
flipper_length_mm	15.78925	11.7826	20.7536