From Flipper to Feather: Exploring Predictive Models for Penguin Body Mass

Jackson Small and Danielle Bodziak

STA4164: Statistical Methods III

Dr. Rong Zhou

December 5, 2023

Abstract

Statistical computing and regression analysis are techniques used to estimate and uncover helpful relationships among variables. Through these means we present to you a study on predicting penguins body mass. In this study, we will be exploring the ideal model to predict body mass in grams. Found on Kaggle, the "Palmer Archipelago Penguin Data" was originally derived from Dr. Kristen Gorman at the Palmer Station Antarctica LTER, who took measurements on three different species (Adelie, Gentoo, and Chinstrap) and islands in Antarctica.

Body mass plays a large role in the survival and reproductive success of penguins, as it directly influences their ability to withstand long fasting periods while caring for eggs. The population dynamics of penguins are intricately linked to the availability of sufficient body mass, making it extremely important to unravel the biological factors contributing to variations in body mass among individuals. This study delved into the specific physical attributes of penguins – namely, bill length, bill depth, and flipper length – to determine their individual and collective impact on body mass. Assumptions of multiple linear regression analysis- existence, independence, linearity, homoscedasticity, normality, leverage and outlier points, and missing value analysis were all examined. The data which verifies these assumptions were statistically computed and analyzed through multiple linear regression in R and linear regression supervised learning in Python. How do physical attributes of penguins along with species classification contribute to the prediction of penguin body mass in grams?

Data Exploration

The penguins dataset includes measurements and variables named as species, island, bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g, and sex. Including missing values the dataset has 344 observations. Shown below are summary statistics to explore frequency distributions of all variables.

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
Min.	32.10	13.10	172.0	2700	----------
1st Qu.	39.23	15.60	190.0	3550	----------
Median	44.45	17.30	197.0	4050	----------
Mean	43.92	17.15	200.9	4202	----------
3rd Qu.	48.50	18.70	213.0	4750	----------
Max.	59.60	21.50	231.0	6300	----------
NA's	2	2	2	2	11

This table shows the four quantitative variables and one qualitative variable sex, showing 11 missing values. Note that species, island, and sex all have 344 observations for each column, but sex. We chose to remove these missing values since there is no way to resample them. For our data analysis we chose our independent variables as bill length, bill depth, and flipper length against our dependent variable, body mass. To understand distributions of our data and visualize it, the histograms and scatterplots are shown. [Graph 1.2] From the histogram plots we can conclude that the independent variables are normally distributed and only showing some skewness in our dependent variable, body mass. [Graph 1.3] Similarly, from our scatterplots, we can confirm that each distribution is linear, although there does seem to be some clustering

between the different species. In order to understand the assumptions of multiple linear regression analysis, we can analyze the residual plot, Q-Q plot, homoscedasticity plot, and the Cook's distance plot. First the existence assumption is met since for every x-value there is a y-value. Next, the independence assumption is met since each value of observations do not affect any other values in the dataset. The linearity assumption is met by looking at the scatterplots and we can see that each independent variable against body mass is linear. Homoscedasticity, according to the plot, shows a random and even scatter demonstrating no concerns and that the variance of y-values are the same for x-values. The normality assumption is also met since the histograms for each independent variable are normally distributed. When analyzing Cook's distance we found that our 294th observation's distance was 0.08 which is a high leverage point compared to the rest of the observations being around 0-0.03. Because of this we decided to remove the 294th observation. This showed a great improvement on our regression plots, therefore showing that there were no more leverage points. For the missing values, we removed all 11 rows which had NA's, therefore our total observations went from 344 to 332. When performing multiple linear regression we must also pay close attention to any variables that can cause Multicollinearity, so that we can accurately estimate the beta coefficients for our model. To measure this we decided to perform VIF (Variance Inflation Factor) Analysis by finding the VIF values. We can remove any that are greater than ten.

	bill_length_mm	bill_depth_mm	flipper_length_mm
VIF value	1.865	1.611	2.673

Since all of the independent variables VIF values are very low we can conclude that there is no concern of multicollinearity for our model. We are now ready to perform statistical computing to select the best model for predicting body mass.

Data Preparation

Although all of the 5 main assumptions were met, we decided to remove all of the NA values and the 294th data point as well to make them fit even better. After doing so, everything remained fairly the same, except for our leverage plot, which showed great improvement. Because of these met assumptions, no transformations were needed to continue the analysis. Although sex of the penguins may have been a deciding factor for body mass, we opted not to include sex as a dummy variable because we were solely focusing on the physical attributes of the penguins. Our full model equation once everything necessary had been removed was Body Mass = β_0 hat + $\beta_1$hat(Bill length) + $\beta_2$hat(Bill depth) + $\beta_3$hat(Flipper length) + ε with an r^2 value of .7647

Modeling

Our univariate models were quite interesting, for the model testing bill length had an r^2 of 0.3608 and bill depth had one of 0.2226. Both of these are quite low and do not imply a strong correlation at all. However, the model including flipper length had an r^2 of 0.763, implying a moderately strong correlation with body mass and we can already assume that flipper length will be the best predictor of body mass in penguins. When we look at the full model comprehensively, we don't find any interaction models. We also see that the only significant variable is flipper length, leading to the reduced model of Body Mass = β_0 hat+ $\beta_3$hat(Flipper length) + ε. For our analysis, we chose to use the Backwards Elimination Approach and the Stepwise Approach. Step one of our backwards elimination approach while removing variables based on p-values was to remove the variable "bill_length_mm" from the model. The next step

was to remove “bill_depth_mm” because it was also not significant. Flipper length was found to still be significant, so we wanted to keep it in the model, resulting in a reduced model of Body Mass = β_0 hat+ $\beta_3$hat(Flipper length) + Ɛ once again. For our stepwise approach, the first step was to add “bill_length_mm” to the model and once we did that, it proved to be significant. The next step was to add “bill_depth_mm” to the model and once again, both independent variables were significant. The third and final step was to add “flipper_length_mm” to the model and this gave us an interesting result, for it caused bill length and bill depth to no longer be significant. This third step also resulted in the same r^2 value of 0.7647.

	0	1	2	3
AIC	4444.499	4297.919	4233.284	3970.164

Looking at these values, we can see that step 3 has the lowest AIC value, indicating a better-fit model. Due to this, we can conclude step 3 is the final step, leading to the reduced model of Body Mass = β_0 hat+ $\beta_3$hat(Flipper length) + Ɛ.

Next we performed training and validation to the dataset in a python environment. Since all the assumptions are met, we can continue with this process. This time for our model we included the species column. In order to proceed with using the species column of the dataset we created three binary coded variables called species_Adelie, species_Gentoo, and species_Chinstrap. Each of these variables are encoded as 1 if they are Adelie, Gentoo, or Chinstrap respectively, and if they are not they are encoded as 0. The other three physical measurements of the penguins and these three new binary variables make up our independent variable against our dependent variable, body mass. Training and testing regression is when the dataset is divided into two subsets called the training and testing sets. In this case 70% of our

data was split into the training set and the other 30% to a testing set. Then using machine learning the model learns patterns and relationships between the features (independent variable) and the target variable. It then takes the training set and tests it against the test set in order to evaluate the model. The full model: Body Mass = β_0 hat+ $\beta_1$hat(Bill length) + $\beta_2$hat(Bill depth) + $\beta_3$hat(Flipper length) + $\beta_4$hat(species Adelie) + $\beta_5$hat(species Gentoo) + $\beta_3$hat(species Chinstrap) + Ɛ. After performing this regression technique our reduced model becomes

$$Body\ Mass = -4019.24 + 44.72(Bill\ length) + 128.80762(Bill\ depth) + 19.6827(Flipper\ Length) - 137.65(species\ Adelie) - 649.16(species\ Gentoo) + 786.81(species\ Chinstrap) + \varepsilon$$

. This model produced an r^2 of 0.8814 and an MSE of 80220.

Conclusion:

After comparing methods we notice that the first three approaches select the same reduced model of $Body\ Mass = \beta_0 hat + \beta_3 hat$ and the Train/Test regression model wof Body Mass = β_0 hat+ $\beta_1$hat(Bill length) + $\beta_2$hat(Bill depth) + $\beta_3$hat(Flipper length) + $\beta_4$hat(species Adelie) + $\beta_5$hat(species Gentoo) + $\beta_3$hat(species Chinstrap) + Ɛ. This summary chart includes important descriptives of our model such as R^2, MSE or AIC, and the number of independent variables.

	Comprehensive	Backwards	Stepwise	Train/Test regression
R^2	0.7647	0.7630	0.7647	0.8814
MSE or AIC	MSE: 154253	MSE: 154440	AIC 3970	MSE: 80220
# of independent variables	1	1	1	6

As we see the Train/test regression cut the MSE in half while also creating increasing r^2 compared to the other models all being the same. When deciding the final model we decided on it being the final model because of its lower MSE score, higher r^2, and its incorporation of species. Our final model that best answers the question, How do physical attributes of penguins along with species classification contribute to the prediction of penguin body mass in grams? is:

$$Body\ Mass = -4019.24 + 44.72(Bill\ length) + 128.80762(Bill\ depth) + 19.6827(Flipper\ Length) - 137.65(species\ Adelie) - 649.16(species\ Gentoo) + 786.81(species\ Chinstrap) + \varepsilon.$$