A Multiple Regression Approach to Predict Bird Masses From Morphometric, Fecundity, and Behavioral Data

Math158 Midterm Project

Ethan Ashby

3/27/2020

Introduction

Birds are a widely dispersed, ecologically-important class with large phenotypic and morphometric diversity (Rosenberg et al 2019). There is no better illustration of bird diversity than body mass, which ranges from the tiny two gram Bee Hummingbird to the Ostrich which is 40,000 times heavier. Understanding bird body mass is critical to understanding avian physiology, ecology, life history, and evolution.

Body mass measurements of birds can be applied in a variety of studies. From an ecological standpoint, body mass was found to be a strong predictor of migratory bird distributions when crossing the Gulf of Mexico (Buler et al 2017). Avian body mass is also critical to their physiology and life history; body mass has been used as a proxy for flight speed and predation risk (Veasy et al 1998). Avian body masses are also important in the study of evolution; a study of bird body masses showed that the evolution of body masses is nonrandom and is constrained by the ability of indivdiuals to turn resources into offspring (Maurer 1998). Body mass prediction using statistical methods is also of interest. Similar regression approaches to the one presented in this project used extant bird morphometric features to estimate body masses and ecological features of extinct flying organisms (Field et al 2013). Thus, study of body mass and developing statistical methods for avian body mass prediction can provide useful physiological and ecological information about extant or extinct birds.

Presented here is a multiple regression approach to predict bird masses from morphometric, fecundity, and behavioral data sourced from textbooks on bird faunas from geographical regions with the best available ecology and behavioral data available. The model presented in this project could be useful to fill the data gap in newly discovered, difficult to capture birds, or extinct birds.

I hypothesize that there will be a strong, linear relationship between body mass and morphometric predictors, characterized by high R^2 and strong predictions. I anticipate that morphometric predictors will dominate the model, and fecundity and behavioral predictors will be ancillary.

In the Methods section, I will discuss the data processing steps (including changing variable encodings, NA and outlier filtering, and variable transformations informed by Box-Cox and diagnostic plots). In the Results section, I will discuss the model selection process, my model of choice, model performance, and application a couple of my favorite birds. Lastly, in the Conclusions section, I will summarize and contextualize my findings and provide possible improvements to this project.

Methods

The dataset used is available through the Ecological Archives (Lislevand et al, 2007). In the original dataset, variables were encoded by sex (e.g. wing length was encoded in three variables 'M_wing', 'F_wing', and 'U_wing' to denote wing lengths of male, female, and unknown sexes). I changed the encoding by collapsing all wing variables into one column called 'wing' and creating another variable called 'sex' to store the sex of each bird. The resulting dataframe had 11,403 rows and 13 columns.

I filtered this dataset for complete cases, so I could exclude rows with missing values for potentially important predictors from the analysis. The dataframe containing complete cases contained 713 entries representing 360 unique species. The resulting dataframe contained the following variables: Family (Numbering of families according to the listing of all bird families in Monroe and Sibley (1997)), Species name (latin), English name (common name), body mass (grams), wing length (mm), tarsus length (mm), bill length (mm), tail length (mm), Clutch Size (average or range midpoint of number of eggs per clutch), Egg_mass (grams), Mating_System (scale of 1-5 ranging from polandry to promiscuity), Display (scale of 1-5 ranging from ground display to aerial display).

A full model was fit, and the Box-Cox method was applied to the full model to inform transformations of the response variable. Box-Cox which suggested a log transformation of the body mass. Plots of all numerical predictors and the transformed response variable were generated (Figure 1), and some morphometric predictors (wing, egg mass, tarsus, bill, tail) were log-transformed to attain linear relationships between predictors and response. I believe that these transformations of the predictor variables were warranted because morphometric measurements should be on the same scale.

The regression assumption of identical, independent samples may be minorly violated in this project. These data represent mean morphological/fecundity/behavioral characteristics of different bird species; these mean traits have different variabilities depending on the sample sizes of birds collected for each species. What's more, the model was fit to the data set's complete cases, which likely overrepresent birds with well-documented physiologies, life histories, and ecologies. The fact that over 10,000 rows were removed by NA filtering may pose challenges to applying this model to little-studied species, as they may have been excluded from the filtering step.

9 birds were removed from the analysis post hoc, as they were outliers with high leverage that may be skewing the model. I verified that these species were morphologically unusual before exclusion: they included 2 species of crane (Brolga and common Crane), 1 species of penguin (Yellow-eyed Penguin), 1 species of hummingbird (Ruby-throated Hummingbird), and 1 species of songbird (Superb Lyrebird).

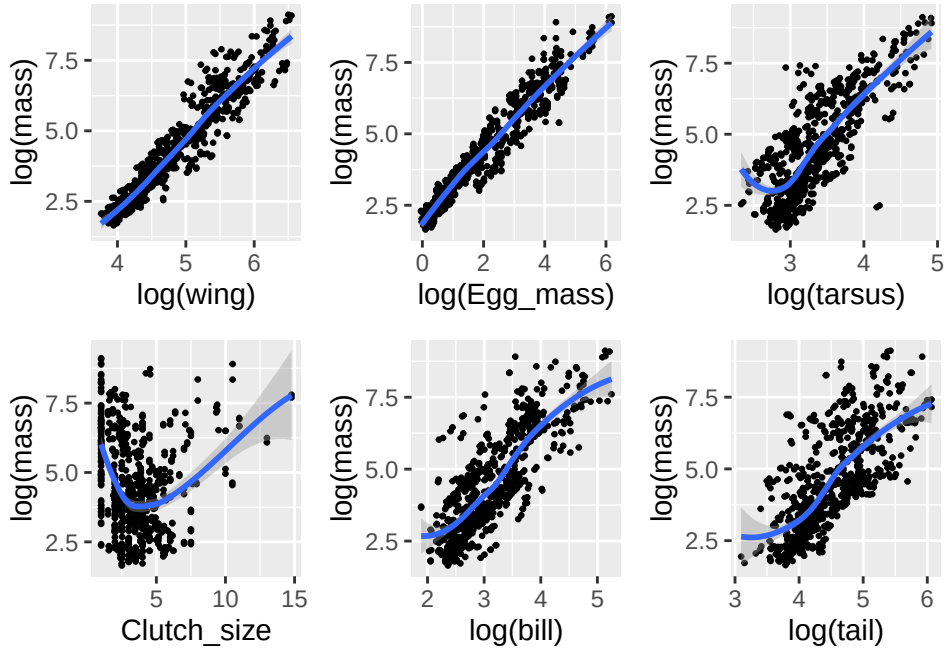

Figure 1: Plots of Transformed and Untransformed Predictors and Log Body Mass

Analysis

I conducted several model selection procedures including comparison of nested models using ANOVA, AIC, and Lasso regression. The nested model ANOVA approach suggested a model with 6 significant predictors (logwing, logegg, logtarsus, Clutch_size, Display, and tail). The AIC approach generated a model with 7 predictors (the 6 from ANOVA plus Mating_system). The LASSO approach suggested a model with 7 predictors (the 7 from AIC).

To include model performance in my selection criteria, I conducted 7-fold Cross Validation (i.e. about 100 observations in each fold) and calculated CV error for each model (Figure 2). As illustrated in the plot, CV error was minimized by the 6 variable model. Given this result, I capped the number of predictors at 6, and excluded the larger models suggested by AIC and LASSO as their higher complexity did not result in improved performance.

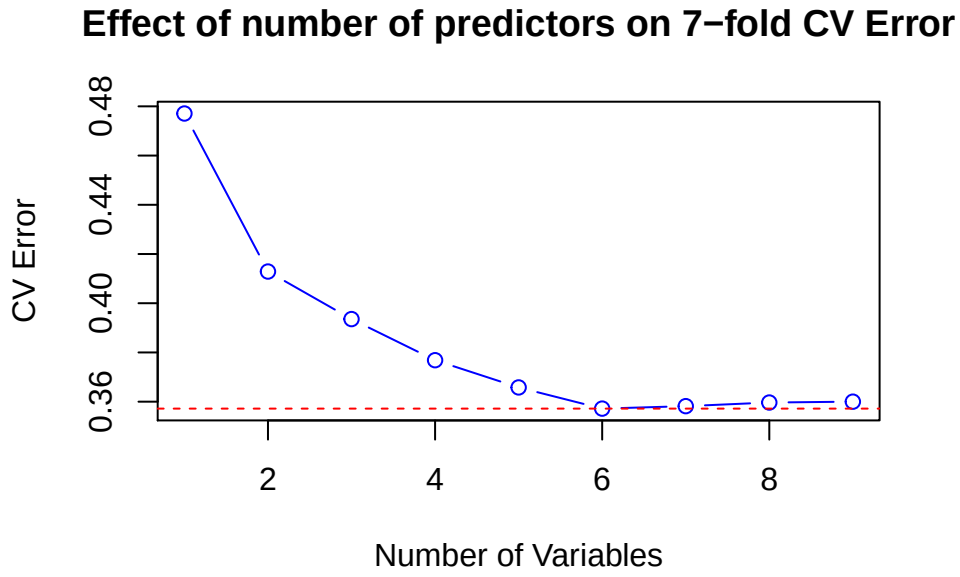

Figure 2: 7-fold CV error of different sized models

The final step of my model selection criteria was performance on independent observations. I revisited the original dataset and filtered for test cases (i.e. cases not included in the training data) that were complete for the 6 variables that my largest model was built on. This generated 715 cases to test my model performance. Then using models built on 1 through 6 predictors, I generated predictions for each case and calculated the squared error. The 3 variable model minimized the mean squared error for this independent test set (Figure 3).

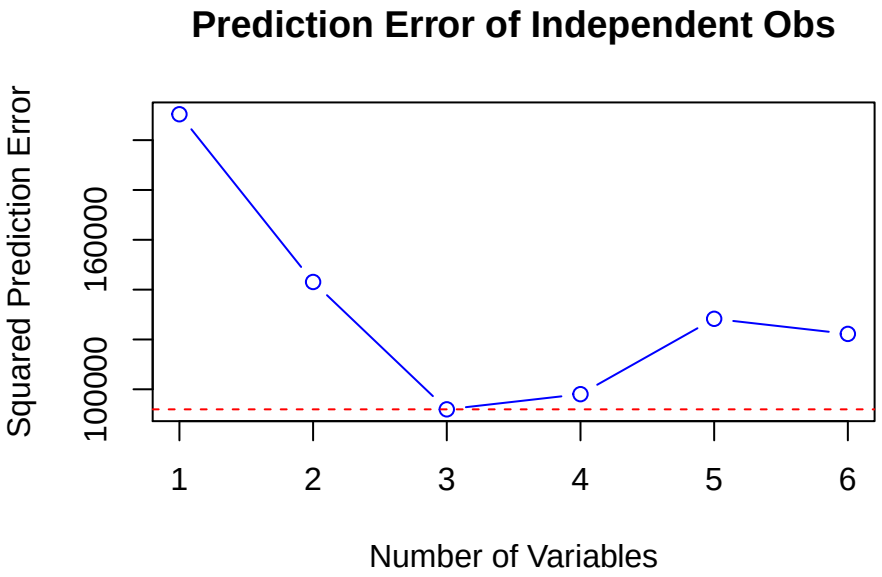

Figure 3: Prediction Error on Independent Test Set of Models Different Size Models

Due to the model's relatively parsimonious structure and it's strong performance in CV error and prediction error, the model that I select as my final model is the 3 variable model (logwing, logegg, and logtarsus). The diagnostic plots suggest the regression assumptions are relatively sound (Figure 6 in Supplementary Plots). The flat line on residuals vs fitted and scale-location plot suggest a good linear relationship, albeit the data is slightly heteroskedastic. Larger fitted values correspond to more variability, which makes sense biologically: lighter passerine birds tend to be more similar in structure, whereas heavier birds can assume completely different morphologies (e.g. albatross and emperor penguin). The Normal Q-Q plot suggests the residuals are approximately normal. The Residuals vs Leverage plot shows that no high leverage points are distorting the regression relationship.

Figure 4: The Greater Prairie Chicken is a heavy, ground-dwelling bird, which is not very well-suited for my model

To show how the model works, I selected a row that corresponded to the Greater Prairie Chicken (Figure 4), a close relative of Cecil Sagehen. Using Prairie Chicken's log wing length, log egg mass, and log tarsus length, the model predicted that the Prairie Chicken would have a weight of 369 grams, when the bird had a weight of 993 grams. This underestimation of the bird's weight is somewhat expected, since that the Prairie Chicken is a ground-dwelling bird, and is thus heavier than it's morphology and fecundity data may suggest.

Figure 5: The Yellow-Rumped Warbler, a smaller passerine and Claremont resident, is more morphometrically standard, and better suited to my model

When the model is applied to a more smaller, more morphometrically standard species, like the Claremont resident Yellow-rumped warbler (Figure 5), our model's prediction improves. Given the morphometric and fecundity data of a Yellow-rumped warbler from our dataset, the model predicted a weight of 15.0 grams, quite close to the true weight of 12.9 grams.

Conclusion

Informed by model selection algorithms and performance metrics, I chose a three variable multiple regression model that predicts log bird species mass using log wing length (mm), log egg size (g), and log tarsus length (mm). The data appeared to fit this linear model quite well, albeit larger fitted values were associated with more variability. The model had a high R^2 (0.95) and performed well in CV and Independent Test Set predictions. The model is also more interpretable compared to more complex models that incorporated other morphometric, fecundity, and behavioral data.

One of this project's major qualitative findings is the importance of fecundity data in predicting bird mass. Intuition might suggest that morphometric data (i.e. data collected from the adult bird itself) would dominate mass prediction. However, log egg mass was found to be the most important variable in the regression analysis (assessed by absolute value of t statistics). This affirms the importance of including fecundity data in predictive models of bird mass and also affirms the importance of collection and dissemination of egg data (like at the Natural History Museum in London).

In terms of improvements that could be made to this analysis, a more comprehensive dataset with fewer missing values would enable more precise predictions for a wider variety of species, including understudied species that may have been excluded from model building in the data filtering step. One could also consider the a model with a phylogenic interaction term, which allows for the fitting of different βs for different taxa, this would also enable more precise predictions for morphologically unusual species like cranes, penguins, and hummingbirds.

My finding that this model preformed better on smaller, standard shaped birds (like the yellow rumped warbler) than morphologically unusual ones (like the Prairie Chicken) suggests that this model is better suited for mass estimation of certain taxa over others. I began an attempt at grouping together these species into higher level taxonomic groups, so that model performance on different taxa could be assessed. My most immediate next step is to conduct this analysis, and confirm my suspicions that this model is tailored to smaller, morphometrically standard passerines.

References

Buler, J. J., Lyon, R. J., Smolinsky, J. A., Jr, T. J. Z., Moore, F. R., Lyon, R. J., & Moore, F. R. (2017). Body mass and wing shape explain variability in broad - scale bird species distributions of migratory passerines along an ecological barrier during stopover. Oecologia, 185(2), 205–212. https://doi.org/10.1007/s00442-017-3936-y

Field, D. J., Lynner, C., Brown, C., & Darroch, S. A. F. (2013). Skeletal Correlates for Body Mass Estimation in Modern and Fossil Flying Birds. PLoS ONE, 8(11), 1–13. https://doi.org/10.1371/journal.pone.0082000

Lislevand T., Figuerola J., Székely, T. (2007). Avian body sizes in relation to fecundity, mating system, display behavior, and resource sharing. Ecology 88, 1605.

Maurer, B. A. (1998). The evolution of body size in birds . I . Evidence for non-random diversification. Evolutionary Ecology, 12(1973), 925–934.

Rosenberg, K. V, Dokter, A. M., Blancher, P. J., Sauer, J. R., Smith, A. C., Smith, P. A., ... Marra, P. P. (2019). Decline of the North American avifauna. Science, 366 (October), 120–124.

Veasy, J. S., Metcalfe, N. B., Houston, D. C. (1998). A reassessment of the effect of body mass upon flight speed and predation risk in birds. Animal Behaviour, 56(4), 883-889.

Supplementary Plots

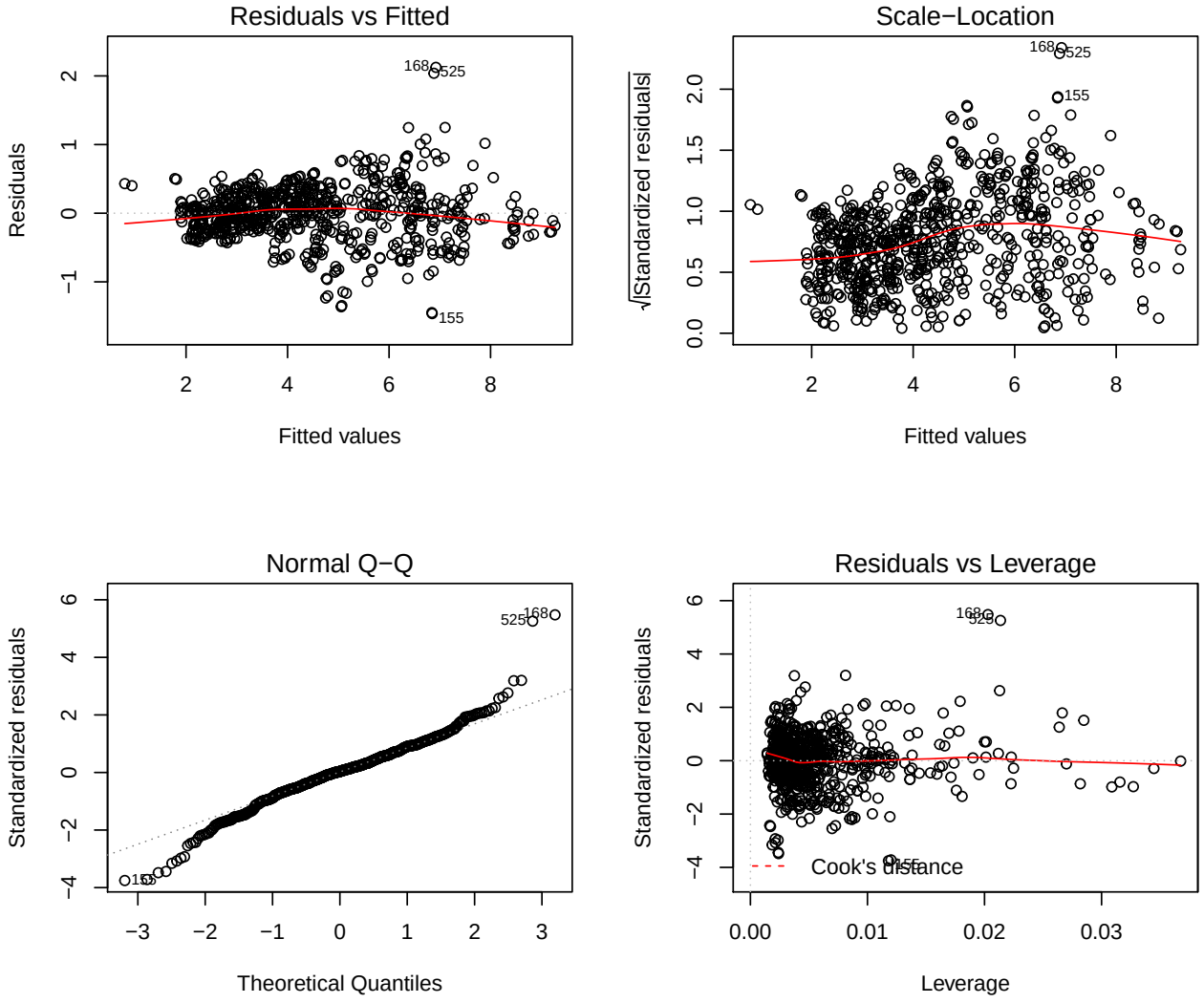

Figure 6: Diagnostic Plots for Final 3-vbl Model