Mid-project check-in Report

Group members: Xingyu Chen, Athan Tsintsilonis, Chengfeng(Austin) Shang

Hypotheses:

Null Hypothesis 1: There is no correlation between average body mass of Aves species and the variance of body mass.
Alternate Hypothesis 1: The average body mass of Aves species is positively correlated to the variances of body mass.

Null Hypothesis 2: The correlation between average body mass and body mass variance is not influenced by body mass differences between male and females.
Alternate Hypothesis 2: The correlation between the average body mass and the body mass variance is under the influence of the body mass differences between male and females

Null Hypothesis 3: The correlation between the average body mass and the body mass variance does not differ with the species longevity.
Alternate Hypothesis 3: the correlation between the average body mass and the body mass variances differ with the species longevity.

Predictions:

1. The average body mass is not dependent to the variance of body mass
2. The average body mass and the variance of body mass are positively correlated
3. The differences in body mass between sexes positively affects the correlation between average body mass and the variance in body mass
4. The species longevity positively affects the correlation between the average body mass and the variance in body mass

Two major revisions were made to our hypothesis. First, we decided to focus on bird (Aves) species instead of mammals (mammalia). This decision was made as more data was available for bird species making it more suitable for analysis. The second revision made was measuring the relationship of body mass and sex, as opposed to body mass and trophic level. This revision was made for convenience sake as sex was already given in the dataset while trophic level of each species would have to be determined manually.

Data Description

The dataset we chose to use contains measurements for 29 life-history parameters across 21322 species of birds, mammals and reptiles. The authors compiled this dataset from various sources, including peer reviewed articles, published books and existing life-history databases. Due to this data being compiled from other sources, it is unlikely that the dataset is biased as each source will likely have a different collection method and therefore it is unlikely any biases from these methods would remain consistent across such a large dataset.

It should be noted that the dataset provided by the authors was organized into four different excel spreadsheets. The two spreadsheets used for our analysis were the main spreadsheet, containing median values for all 29 life-history parameters of each species, and the range count spreadsheet, containing maximum and minimum values for each parameter from the main spreadsheet. While not used, the author's also provided a reference spreadsheet containing references for each datapoint as well as a sparse spreadsheet containing the raw data used to construct the main spreadsheet.

Our dataset was manipulated in 4 major steps. To start off, all bird species were extracted from the main dataset by using the filter() function to filter for all species from the Aves class. Secondly, our variables of interest were gathered using the select() function. These variables include adult body mass, species longevity, female body mass and male body mass. Minimum and maximum values for these variables were also collected from the range count dataset using the same method. Next all rows containing the value -999 were removed from the dataset using the filter function. This removed all invalid entries from the dataset because a -999 value was given to all unknown values by the authors. Finally, all extracted data was combined into a single dataframe using the cbind() function. The final data frame contains 10 life-history parameters of 5026 Aves species.
Next, we calculated the variances of body mass for different species, and for different sexes, and assigned those values into newly created columns.

Data Analysis Plan
We plan to perform a simple linear regression model with body mass variances as predictor and average body mass as response, this model is aimed to test the Hypothesis 1.
We plan to perform a multivariate linear regression model, with body mass variance,sexual body mass differences as predictors and the average body mass as the response, this model is aimed to test the hypothesis 2.
We plan to perform a multivariate linear regression model, with body mass variance,sexual body mass differences and species longevity as predictors and the average body mass as the response, this model is aimed to test the hypothesis 3.
To perform a linear regression model, we first need to make sure there is a linear relationship between the predictors and response, and we plan to verify it with a simple scatterplot.
For multivariate linear regression model, we have to meet 2 assumptions:

1. The predictors are all normally distributed.
2. The variances of 3 predictors are homogenized.

To test for these 2 assumptions, we will use Residual vs. Fitted plot and Normal Q-Q plot.

We may also perform a PCA on our data. The variables includes average body mass, average male body mass, average female body mass, body mass variances, sexual body difference and longevity.
For PCA, we have to meet 3 assumptions:

1. there are linear relationships among variables.
2. all variables are quantitative data
3. needs more rows than columns in the datasheet.

Assumptions 2 & 3 are already met, since all variables are numerical and there are more rows (1395) than columns (21)

We still need to check for the linearity using scatterplot.
Data Source:

P. Myhrvold, Nathan; Baldridge, Elita; Chan, Benjamin; Sivam, Dhileep; L. Freeman, Daniel; Ernest, S. K. Morgan (2016): An amniote life-history database to perform comparative analyses with birds, mammals, and reptiles. Wiley. Collection.
https://doi.org/10.6084/m9.figshare.c.3308127.v1