Topic 1: Data transformation to a normal distribution to find how wing length differs among hawk species.

STA 106 Project 2 Topic 1

3/6/2024

A. INTRODUCTION

In this statistical report, we explore the variance in wing length among three species of hawks: Cooper's (CH), Red-tailed (RT), and Sharp-shinned (SS). The primary objective is to understand how the wing length, in millimeters, differs across these species. For our analysis, the distribution of the data has to be normal and the variance has to be constant to meet one of the assumptions of our Analysis of Variance (ANOVA) model. We will attempt to find the best combination of outlier removal and transformation to meet this assumption.

B. INITIAL DATA PLOTTING

For this study, we will fit the data using the single-factor ANOVA (SFA) group means model which assesses the relationship between the hawk's average primary wing feather length in mm (μ_i) vs the three species (where i = CH, RT, SS). The model is as follows where Y_{ij} is the unknown population's mean length of feather wing length for each hawk species plus individual error (ϵ_{ij}):

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

For this model, we must satisfy assumptions of normality, constant variance, and group mean independence.

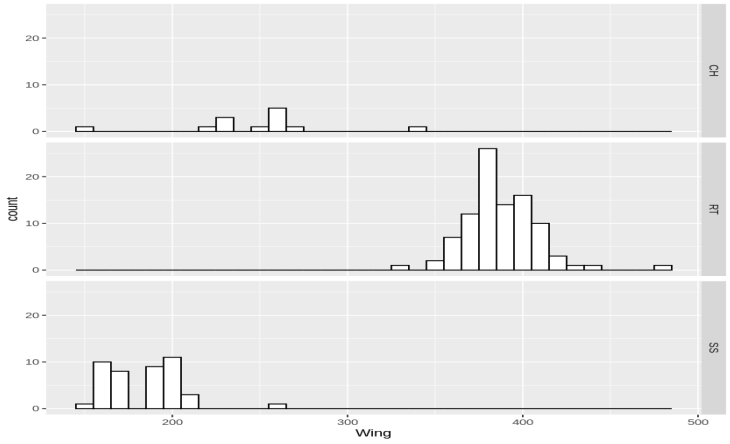
To determine which alpha (the threshold) to use for diagnostics, we need to assess whether we want to minimize the chance of a Type I Error or a Type II Error.

- Type I Error: When you reject the null hypothesis when in reality it is true. In this case, this represents the chance we conclude the data violates our ANOVA assumptions when in reality it satisfies our ANOVA assumptions.
- Type II Error: When you accept the null hypothesis when in reality it is false. In this case, this represents the chance we conclude the data satisfies our ANOVA assumptions when in reality it violates our ANOVA assumptions.

For determining normality and constant variance, we want to minimize our probability of incorrect assumptions, so a Type II error is worse than a Type I error. As a result, we want to maximize alpha, so we will use 0.1 as our alpha value.

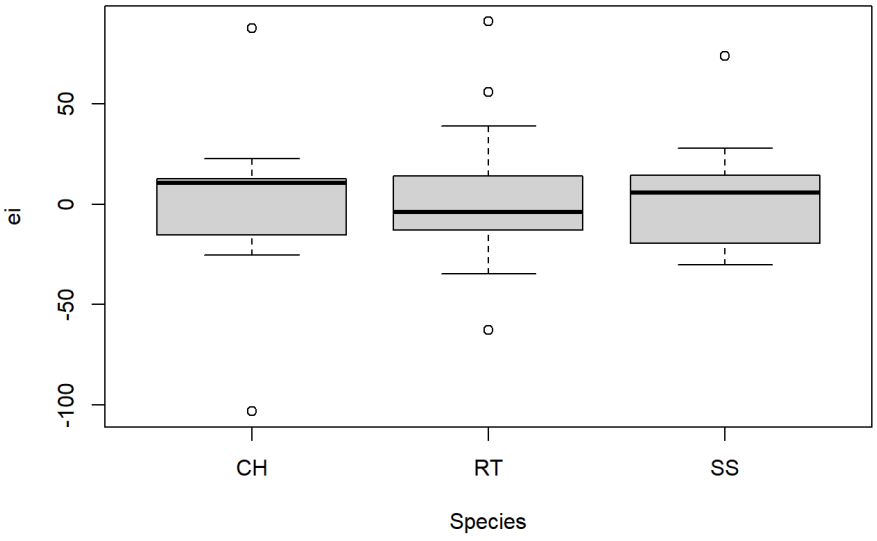
I. Original Data Overview

Figure 1.1.1 - Histogram of Wing Feather Length by Group



We will first assess normality and check if our data does not violate this assumption. We can briefly observe our data using histograms of wing feather length by group and looking at the curve. Figure 1.1.1 shows histograms that show the distribution of hawk feather length by species. It suggests that the normality assumption may be violated for the Cooper's and Sharp-Shinned Hawks. However, for the Red-tailed Hawks, the data looks approximately normal (the shape is symmetrical bell curve).

Figure 1.1.2 - Box Plot of Different Species

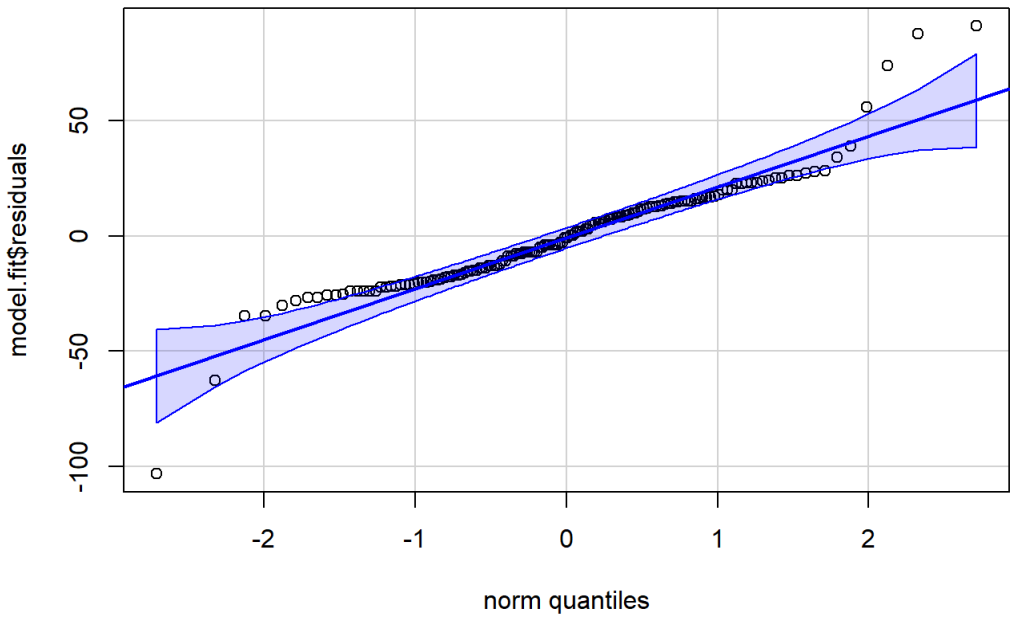


We will now assess outliers and variance. Figure 1.1.2 shows that we have outliers. These outliers can affect our normality distribution and variance of our groups. While our variances

appear approximately constant in this figure, this may not be the case due to the outliers. We cannot conclude without doing formal testing.

II. Normality of data

Figure 1.2 - QQ Plot of our original data

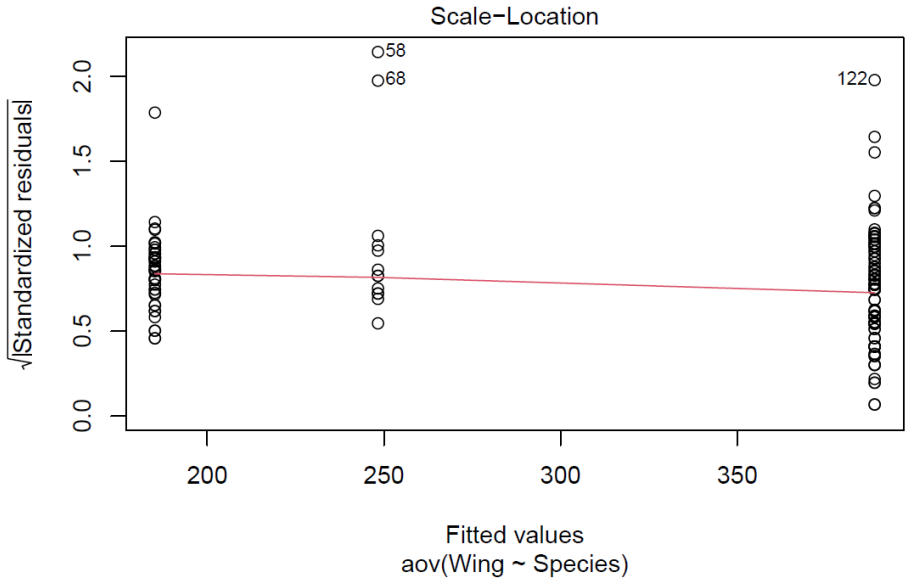


We can also speculate our normality assumption by fitting our model into a QQ plot that compares the quantiles of residuals against the quantiles of a normal theoretical distribution. Figure 1.2 shows the QQ plot as a “non-normal” plot with small and large outliers. However, this is not a conclusive test, so we will run the Shapiro-Wilks test as a formal test.

We use the Shapiro-Wilks test in R to quantitatively assess the normality of our model. Our null hypothesis is that the residuals of our data are normally distributed. Our alternative hypothesis is that they are not normally distributed. We got a p-value of less than 0.001, and this value is less than a significance value of 0.1. Therefore, we reject the null hypothesis and conclude that the values of the residuals are non-normal. This test allows us to claim that the original data is not normally distributed.

III. Constant Variance

Figure 1.3 - Error Plot of our Original Data



In figure 1.3, the variances between groups also appear equal. However, this can easily not be the case as the outliers are heavily affecting the variance for two groups. It is clear that if we remove the outliers, the variances will no longer appear equal.

We can run the Brown-Forsythe test in R to claim to test the assumption for constant variance. Our null hypothesis is that all group variances are equal while our alternative is that at least one group variance is not equal. We calculated a p-value of 0.0991 which is less than a significance value of 0.1. We reject the null and we can conclude that not all group variances are equal. Therefore, the constant variance assumption is also not met for our model.

C. TRANSFORMATION OF DATA

We will consider all possible box cox transformations and outlier removal techniques to better fit our model.

I. Transformation Possibilities

We will consider the following box-cox transformations:

1. PPCC
2. Shapiro-Wilks
3. Log-Likelihood

II. Outlier Removal Possibilities

We will consider the following outlier removal techniques:

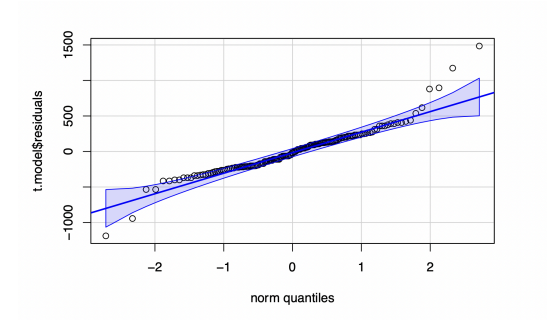
1. Removing Box Plot Outliers
2. Outlier Removal via Studentized Residuals

Note that we cannot consider removing outliers via Semi-Studentized residuals since the variances of the original data are not constant. Also note that when we combined outlier removal with transformation, we performed outlier removal before the transformation.

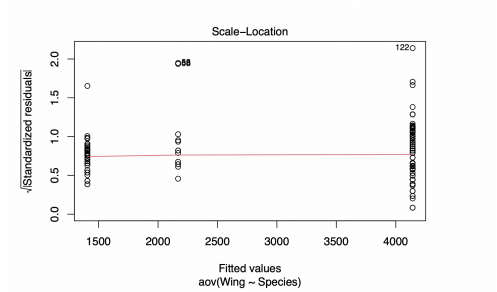
III. Resulting QQ Plots and P-value Tables

Plots for no outlier removal with PPCC transformation:

QQ Plot

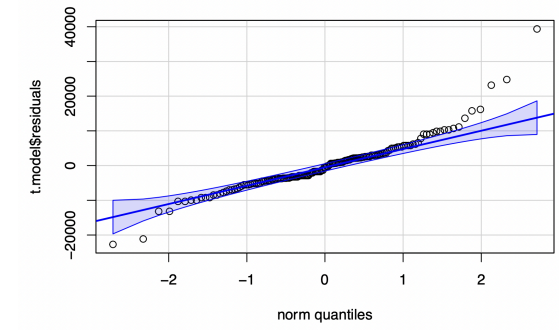


Error Plot

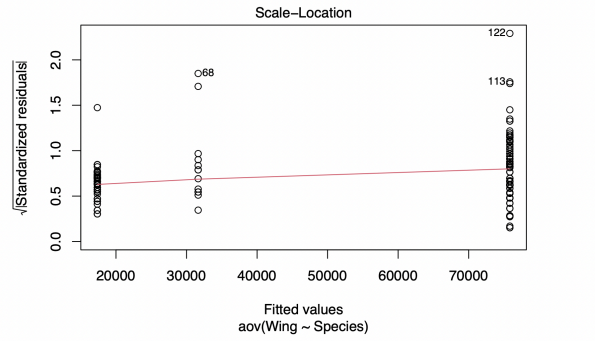


Plots for No outlier removal, log likelihood transformation:

QQ Plot

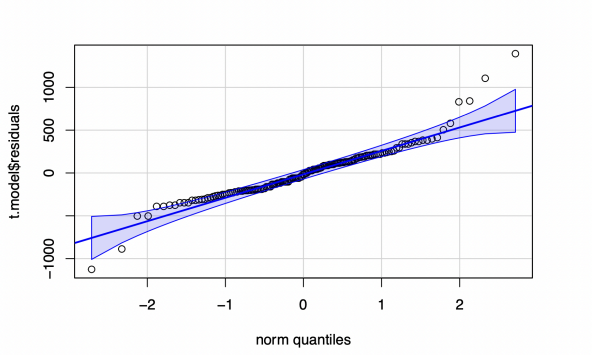


Error Plot

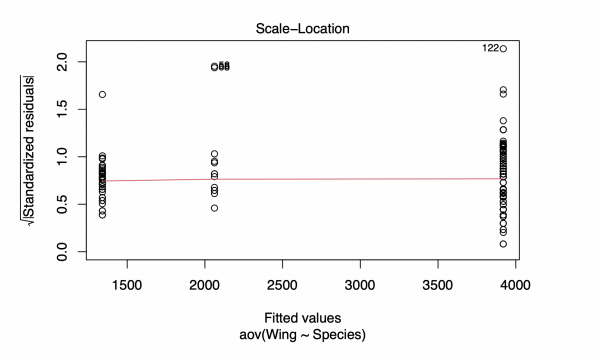


Plots for Not outlier removal, SW transformation:

QQ Plot

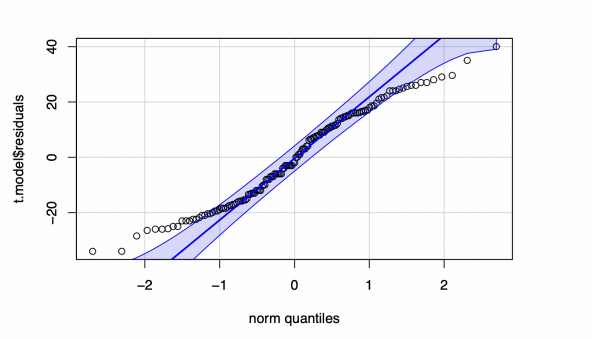


Error Plot

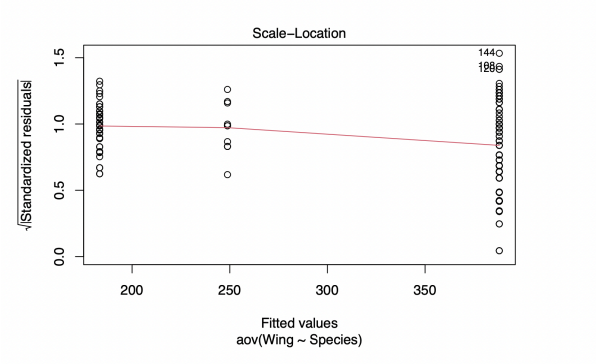


Plots for Removing outliers using boxplot:

QQ Plot

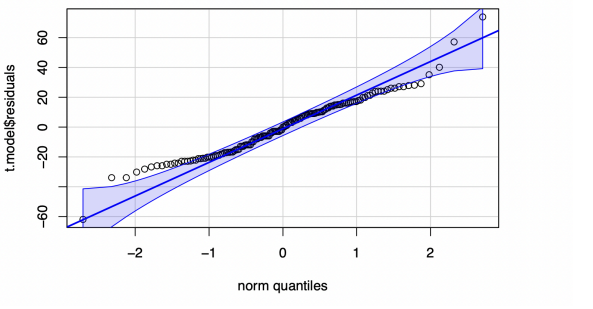


Error Plot

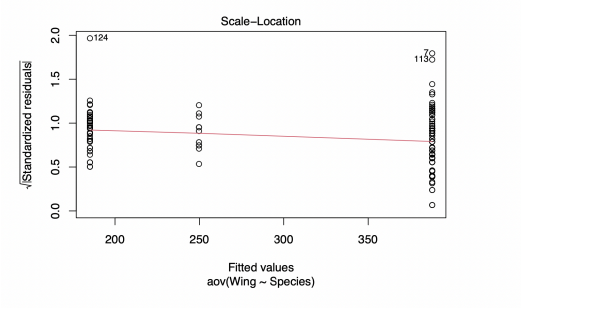


Plots for Removing Outliers via Studentized Residuals:

QQ Plot

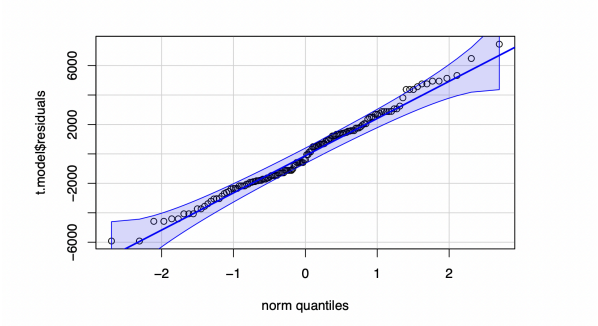


Error Plot

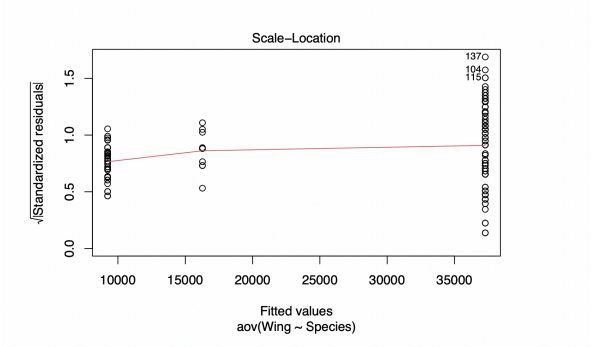


Removing outliers using Box Plot, using PPCC Transformation:

QQ Plot

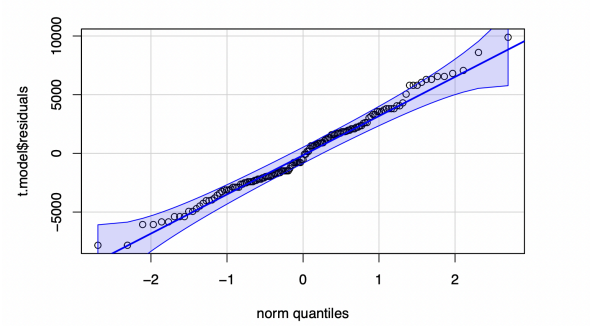


Error Plot

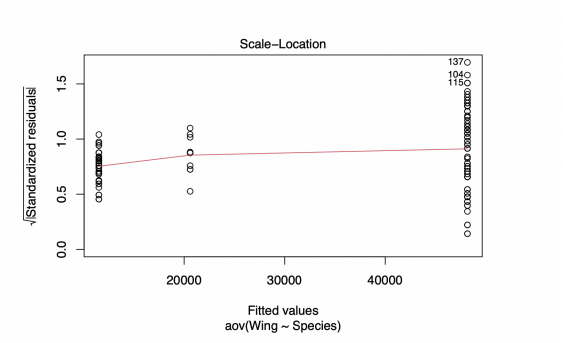


Remove outliers Box Plot, SW Transformation:

QQ Plot

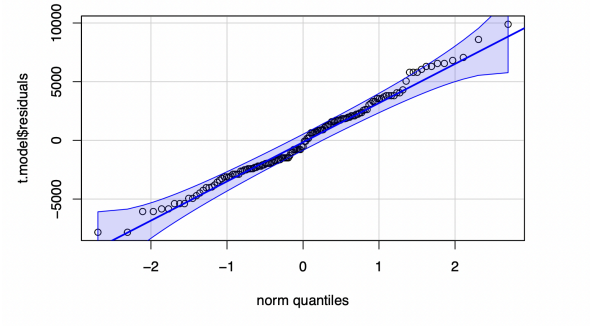


Error Plot

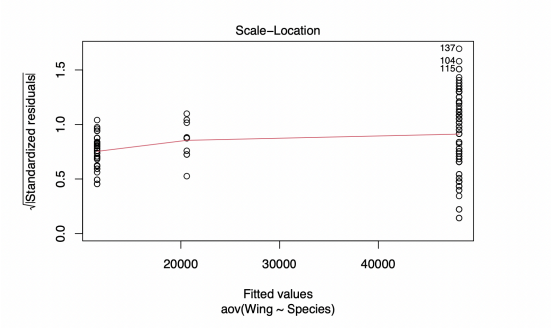


Remove outliers using Box Plot, Log Likelihood Transformation:

QQ Plot

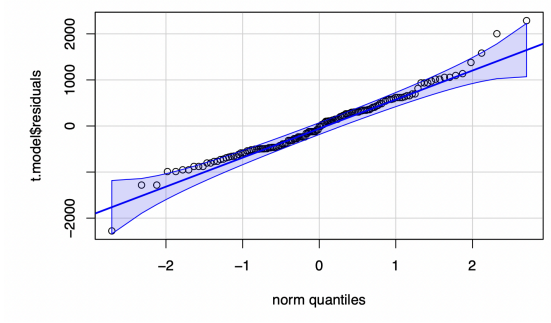


Error Plot

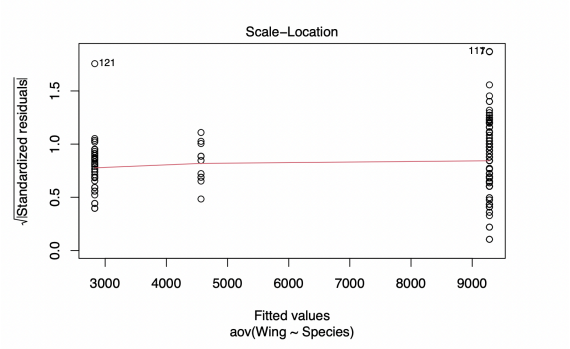


Plots for removing outliers via Studentized Residuals and PPCC Transformation:

QQ Plot

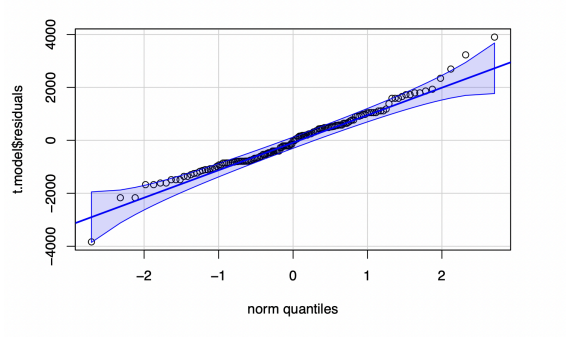


Error Plot

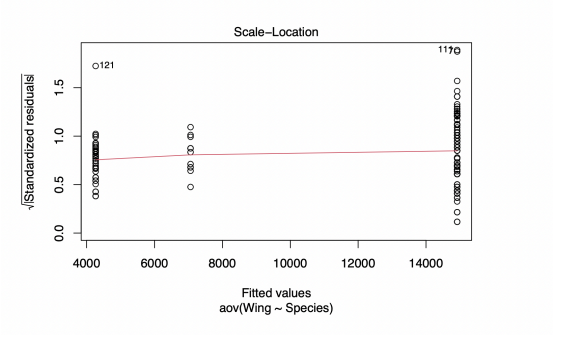


Plots for Removing outliers via Studentized Residuals, SW Transformation:

QQ Plot

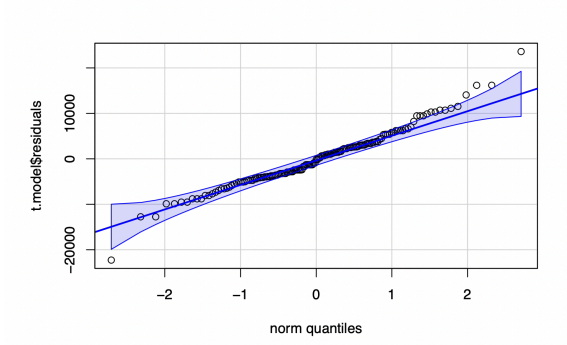


Error Plot

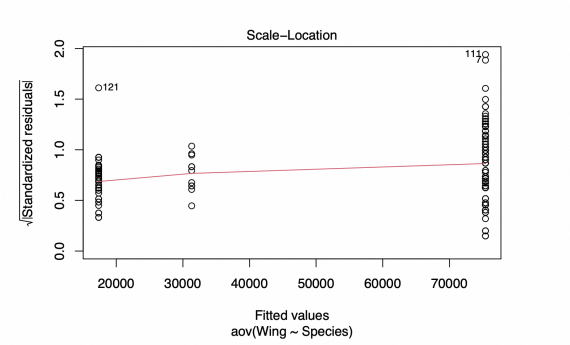


Plots for Removing outliers via Studentized Residuals, Log likelihood transformation:

QQ Plot



Error Plot



P value table for transformations with no outlier removal:

	PPCC	Log Likelihood	SW
Test for normality (SW Test)	0	0	0
Test for constant variance (BF Test)	0.2001	0.0211	0.2035

P value table for outlier removal with no transformation:

	Boxplot	Studentized Residuals
Test for normality (SW Test)	0.0029	0.0040
Test for constant variance (BF Test)	0.3571	0.3769

P value table for outlier removal and transformation:

	PPCC	SW	Log Likelihood
Box Plot Outlier Removal normality test (SW Test)	0.1952	0.1981	0.1848
Box Plot Outlier Removal const variance (BF Test)	0.0103	0.0058	0.0021
Studentized Residuals Outlier Removal Normality Test (SW Test)	0.0116	0.0118	0.0068
Studentized Residuals Outlier Removal const variance (BF Test)	0.1778	0.0893	0.0052

BEST MODEL

Our best models are PPCC transformation with Box Plot Outlier Removal if normality is more important, or PPCC transformation with Studentized Residuals Outlier Removal if variance is more important. While these yielded the highest minimum p-values for both normality test and constant variance test, these are unfortunately still below our threshold of alpha of 0.1, which means we cannot conclude that these meet our ANOVA requirements.

DISCUSSION

Our strategy for transformation was to consider every combination of possible outlier removal and transformation techniques. We employed outlier techniques such as utilizing box plots and semi-studentized residuals in order to ensure a more normal fit and to improve the robustness of the analysis so that it becomes less sensitive to outliers.

If we were to transform the data without outlier removal, the p-value table shows that all transformations (PPCC, SW, or Log Likelihood) failed to achieve normality. This is shown by p-values of 0 in the tests for normality which is significantly below the alpha level of 0.1. These are not worth considering.

If we were to use outlier removal techniques (boxplot or Studentized Residuals) and no transformation, there was a notable improvement in achieving data normality, as evidenced by higher p-values 0.0029 and 0.0040, respectively in the normality tests, but these values are still far below the our threshold of 0.1, so the distribution is still not normal. These are also not worth considering.

If we were to use outlier removal techniques followed by a transformation, we find that PPCC transformation combined with any outlier removal were our best models. There was an improvement in normality for PPCC and Box Plot Outlier Removal, where the p-value is 0.1952. Unfortunately, this still does not meet our criteria since the p-value for constant variance test is 0.0103, well below our alpha value of 0.1. Another interesting result is that if we use PPCC and Studentized Residuals Outlier Removal, we satisfy the constant variance assumption, where the p-value is 0.1778. However, we still don't meet the normality assumption since the p-value is 0.0116, well below our alpha of 0.1. Any other combination of outlier removal and transformation is not worth considering as it only yielded worse results.

In conclusion, there is no good combination of transformed variables, since for any combination of outlier removal and transformation, we will result with either non-normal data or unequal variance or both, considering our alpha value is 0.1. However, if we were to allow a more generous alpha value of 0.01, then we can use PPCC transformation and Box Plot Outlier

Removal if normality is more important, or PPCC transformation and Studentized Residuals Outlier removal if constant variance is more important. In both cases, we would satisfy our ANOVA assumptions if we allowed a more generous alpha value of 0.01. But, this low of an alpha value isn't typical for these tests, as the typical alpha values are 0.05 or 0.1.

One of the downsides of transformation is the challenge of complex data interpretation. After transformations, the data gets harder to interpret because the differences in terms of the original units become less straightforward within the context of the study. Without transformation, interpreting differences in wing lengths between hawk species is as stated. For example, if the average wing length of a Cooper's hawk is 250 mm and the average wing length of a Red-tailed hawk is 350 mm, it can be concluded that Red-tailed hawks have 100 mm longer wings than Cooper's hawks. However, when we apply a transformation to meet the ANOVA assumptions, interpreting differences becomes more complicated. For example, the new data no longer represents wing lengths in a way you can compare them. A difference of 0.2 in the transformed scale does not translate to a difference in wing length (in mm) without reconverting it back to the original scale.