Project 2 Outliers and Transformation

Andrew Jowe

A.1 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.1 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 μ_i vs the three species (where i = CH, RT, SS). This model follows where Y_{ij} is the unknown population's mean length of feather wing length for each hawk species plus individual error is $Y_{ij} = \mu_i + \epsilon_{ij}$.

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

B.1.1 Determining α for our testing

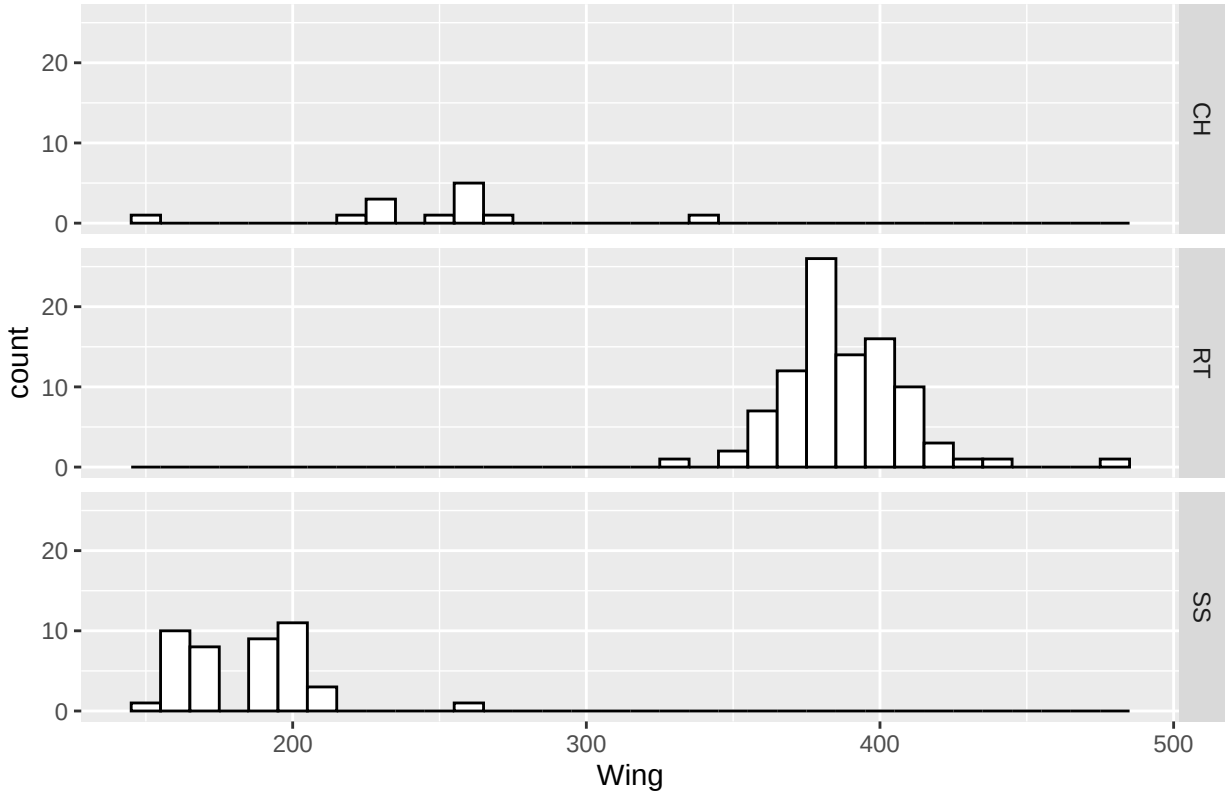
To determine which α to use for diagnostics, we need to assess whether we want to minimize the chance of a Type I Error or a Type II Error.

- Type I Error: When you reject H_0 when in reality H_0 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 H_0 when in reality H_0 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 α, so we will use $\alpha = 0.1$ as our threshold.

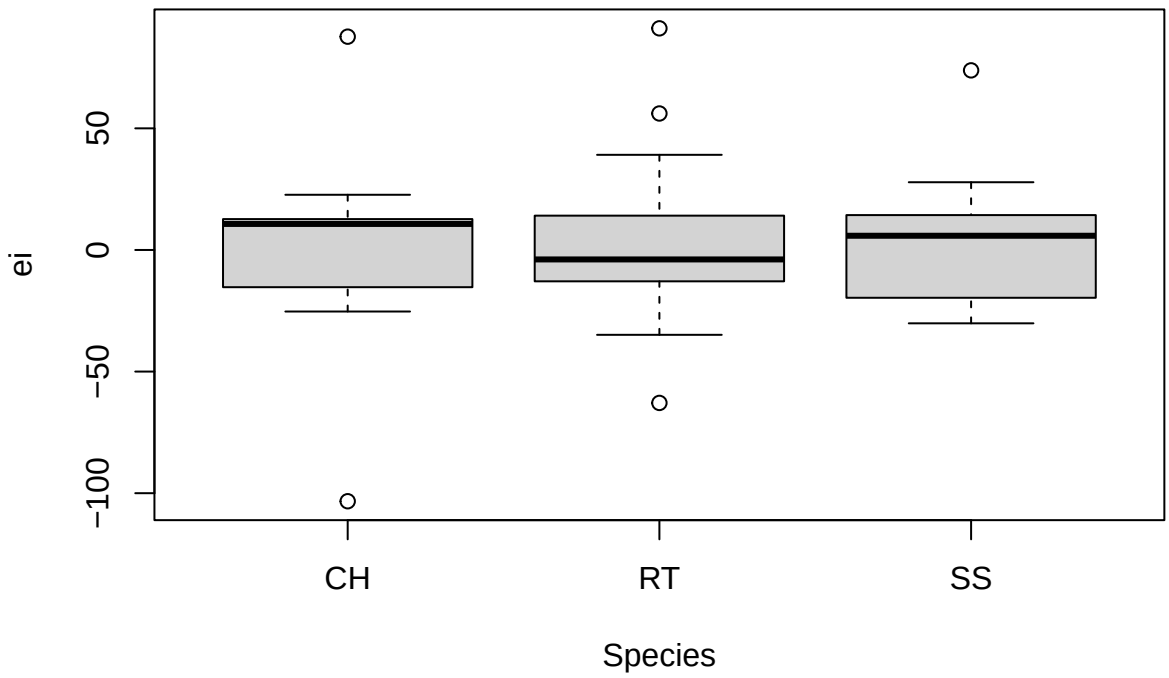
B.1.2 Original Data Overview

Figure 1.1.1 Histogram of Wing Feather Length by Group



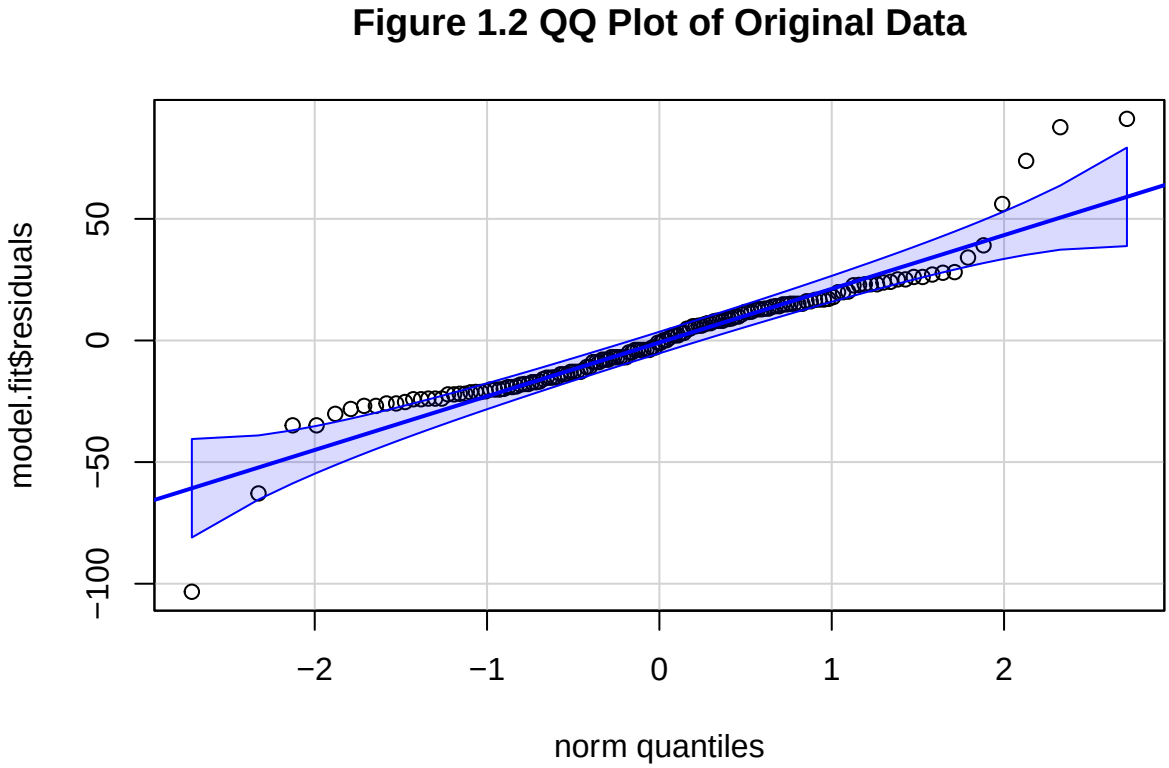
We will first assess normality and check if our data does not violate this assumption. We can briefly observe our data using histograms of wing feather length by group and looking at the curve. Figure 1.1.1 shows histograms that show the distribution of hawk feather length by species. It suggests that 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.

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.

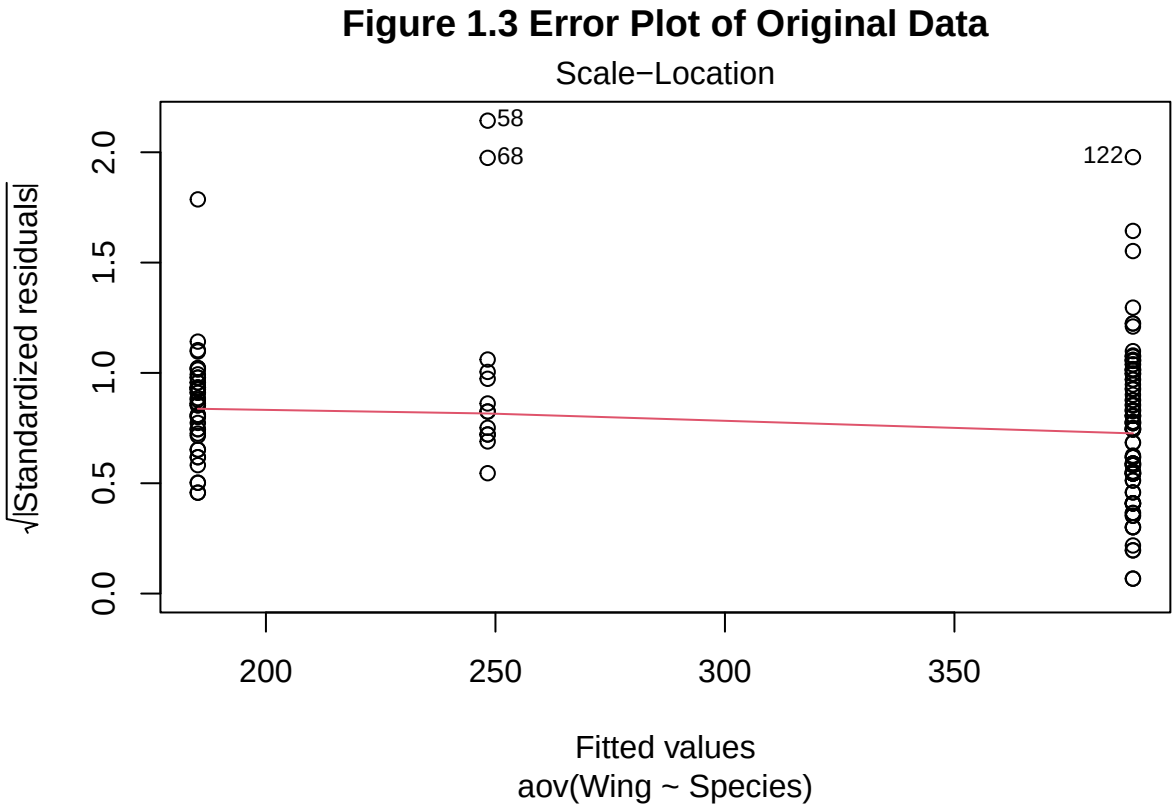
B.1.2 Normality of data



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. 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 0, 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.

B.1.3 Constant Variance



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 Outlier Removal and Transformation

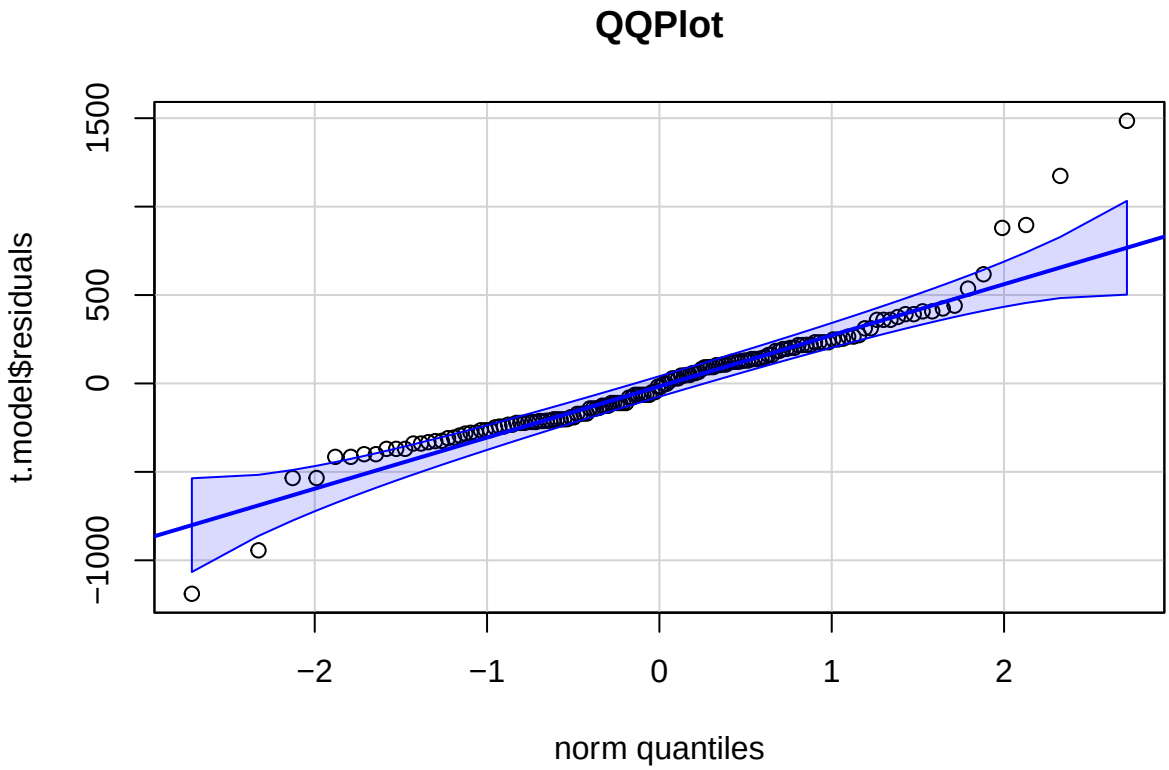
We will consider all possible box cox transformations and outlier removal techniques to better fit our model.

C.1 Transformation Possibilities

We will consider the following box-cox transformations:

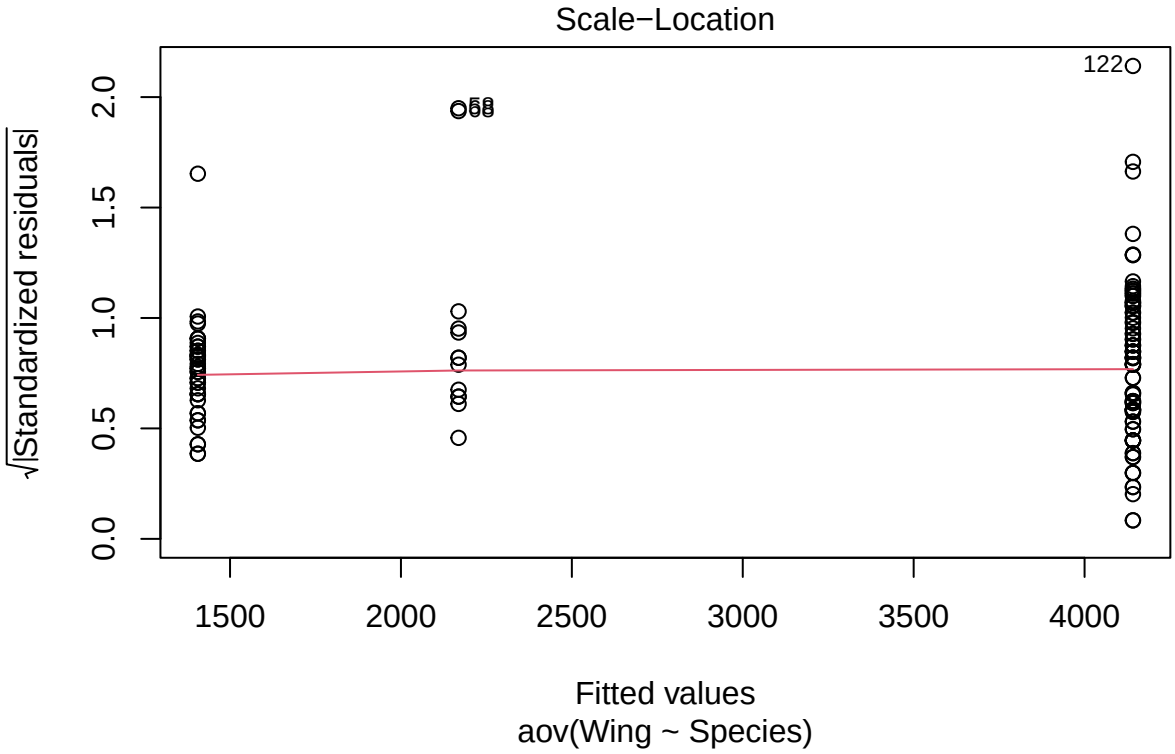
1. PPCC
2. Shapiro-Wilks
3. Log-Likelihood

C.1.1 No outlier removal, PPCC Transformation



Test for normality

Plot variances

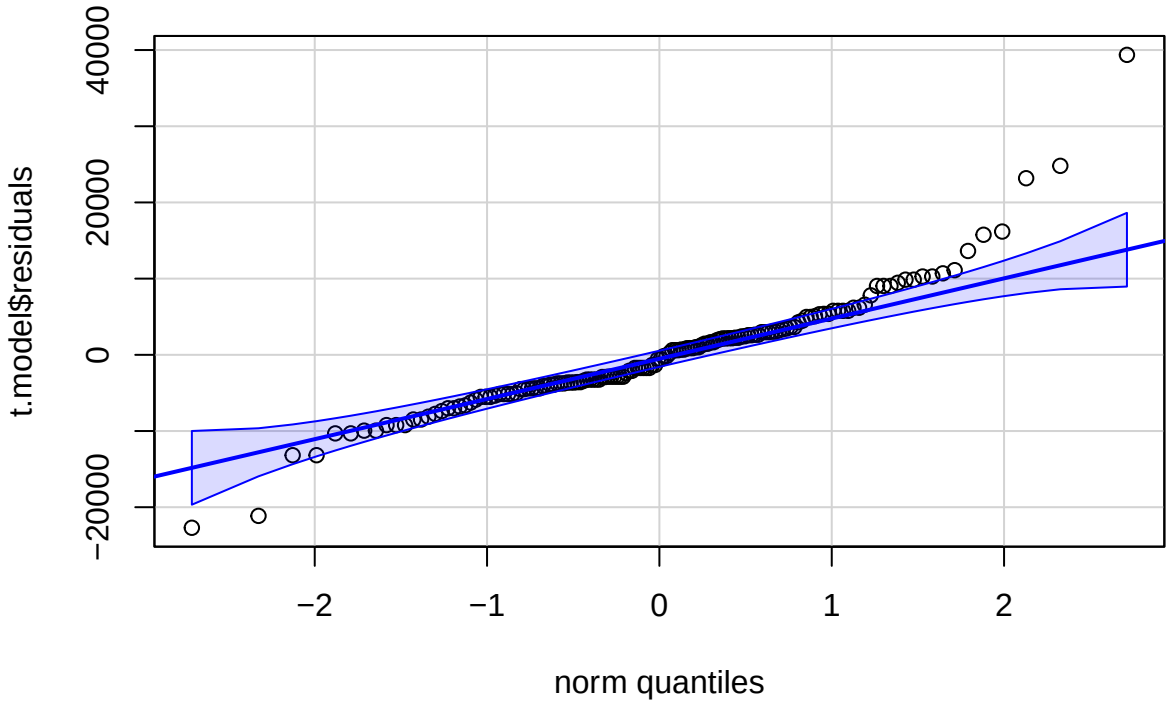


Test for const variance

No outlier removal, Log Likelihood Transformation

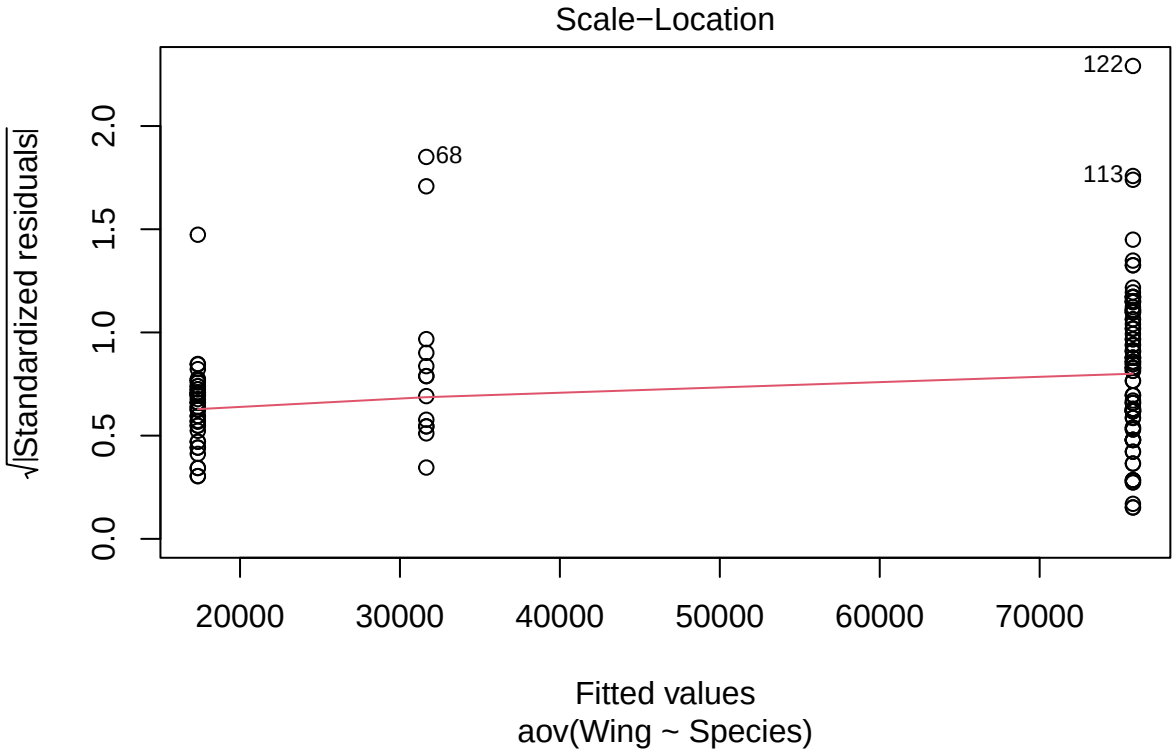
```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## Species       2 1.088e+11 5.440e+10   956.7 <2e-16 ***
## Residuals   147 8.358e+09 5.686e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Get QQ Plot



Test for normality

Plot variances



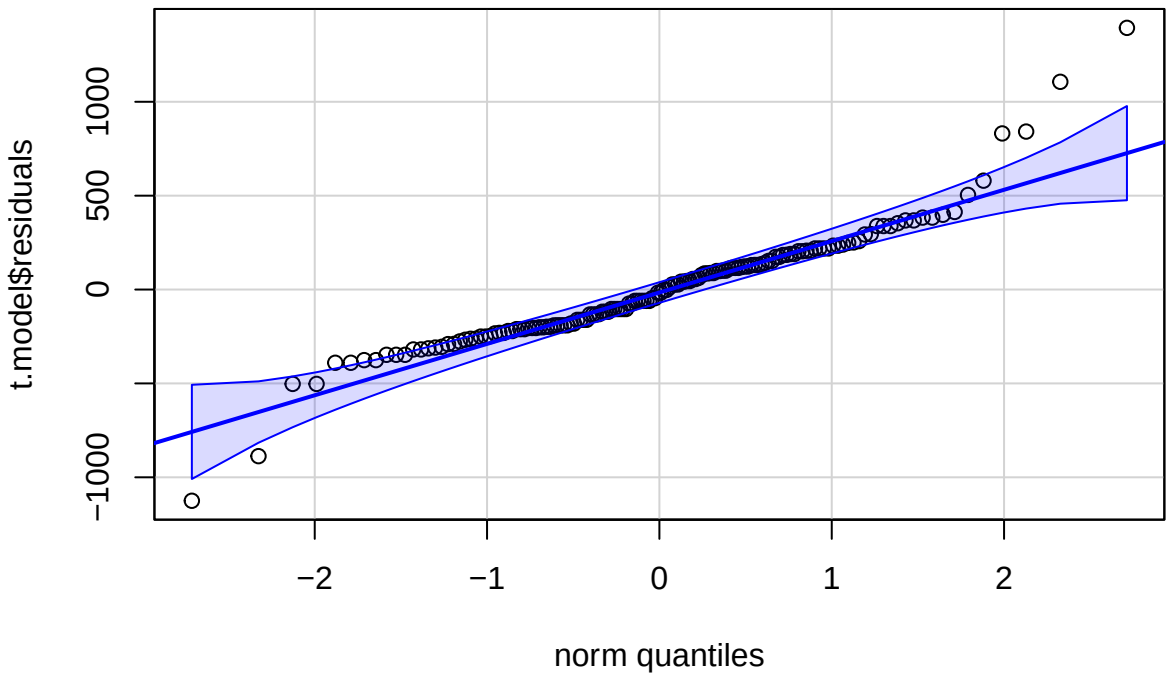
Test for const variance

	Df	F value	Pr(>F)
group	2	3.962041	0.0210903
	147	NA	NA

No outlier removal, SW Transformation

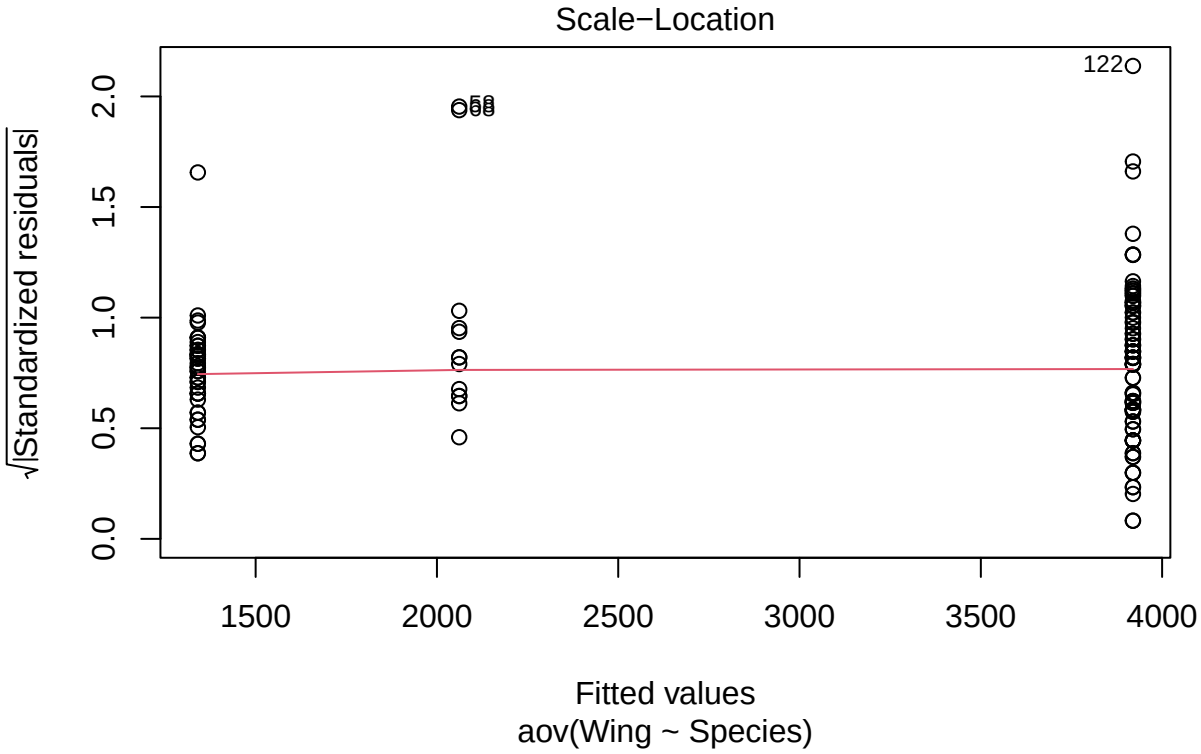
```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## Species       2 209280775 104640388    1113 <2e-16 ***
## Residuals   147  13820960     94020
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Get QQ Plot



Test for normality

Plot variances



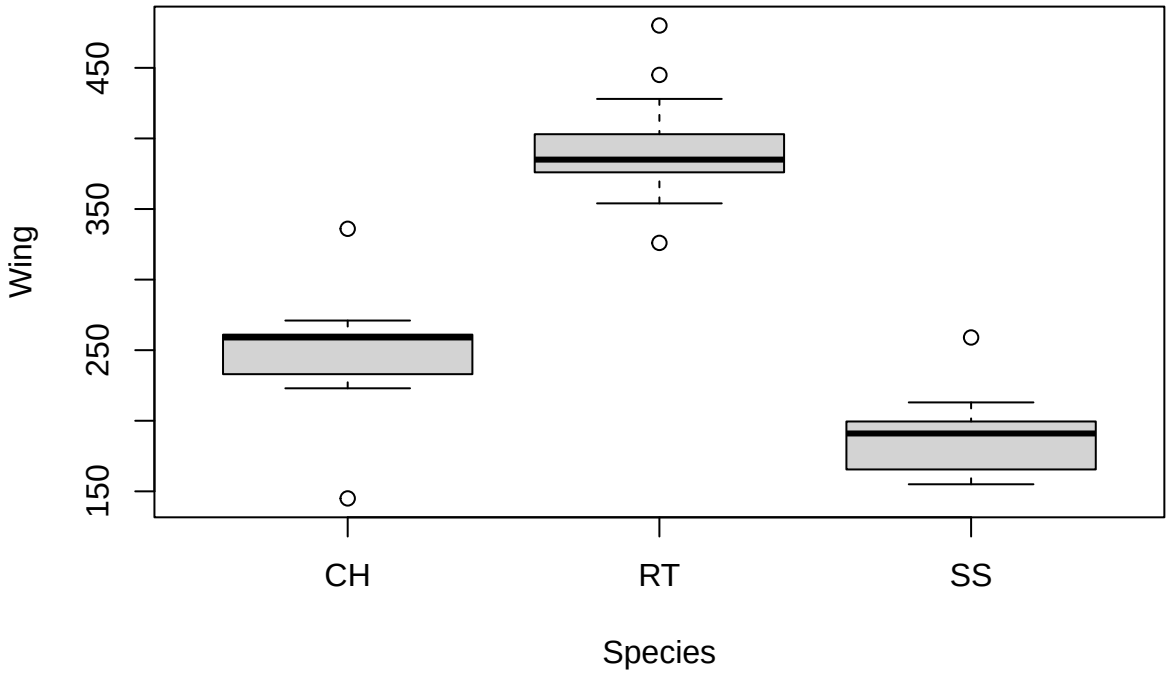
Test for const variance

	Df	F value	Pr(>F)
group	2	1.609378	0.2035159
	147	NA	NA

Possible outlier removal techniques

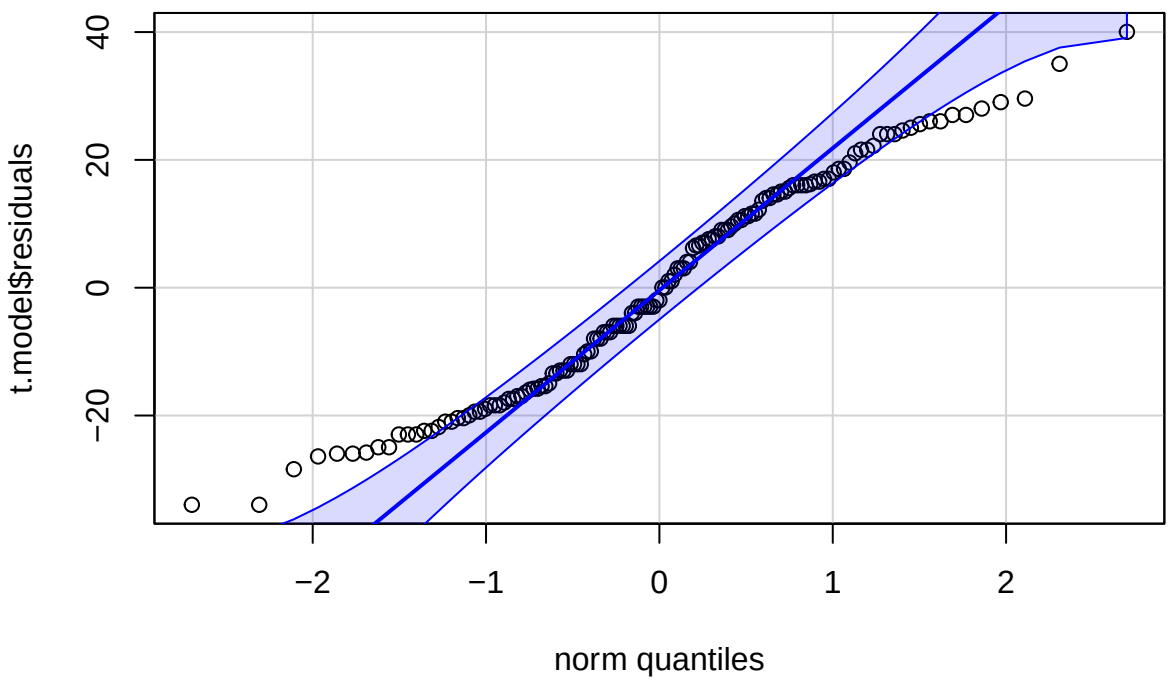
1. Outlier removal via box plot
2. Semi-Studentized Residuals: we can use this since we have the assumption that our variance is constant from our original test. We don't need to do studentized residuals since this is a more robust replacement.

Removing outliers via box plot (1)




```
##              Df  Sum Sq Mean Sq F value Pr(>F)
## Species       2 1254230  627115    2141 <2e-16 ***
## Residuals   140   41013     293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

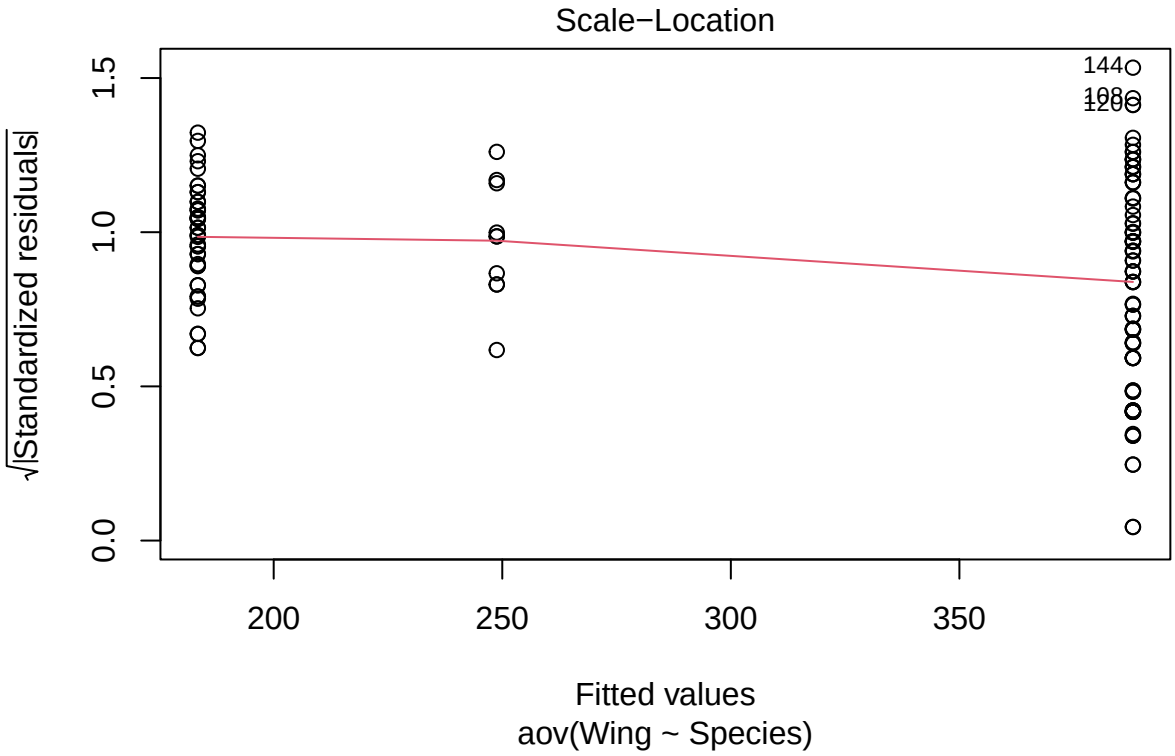
Get QQ Plot



Test for normality

```
##
##  Shapiro-Wilk normality test
##
## data:  t.model$residuals
## W = 0.96964, p-value = 0.002881
```

Plot variances



Test for const variance

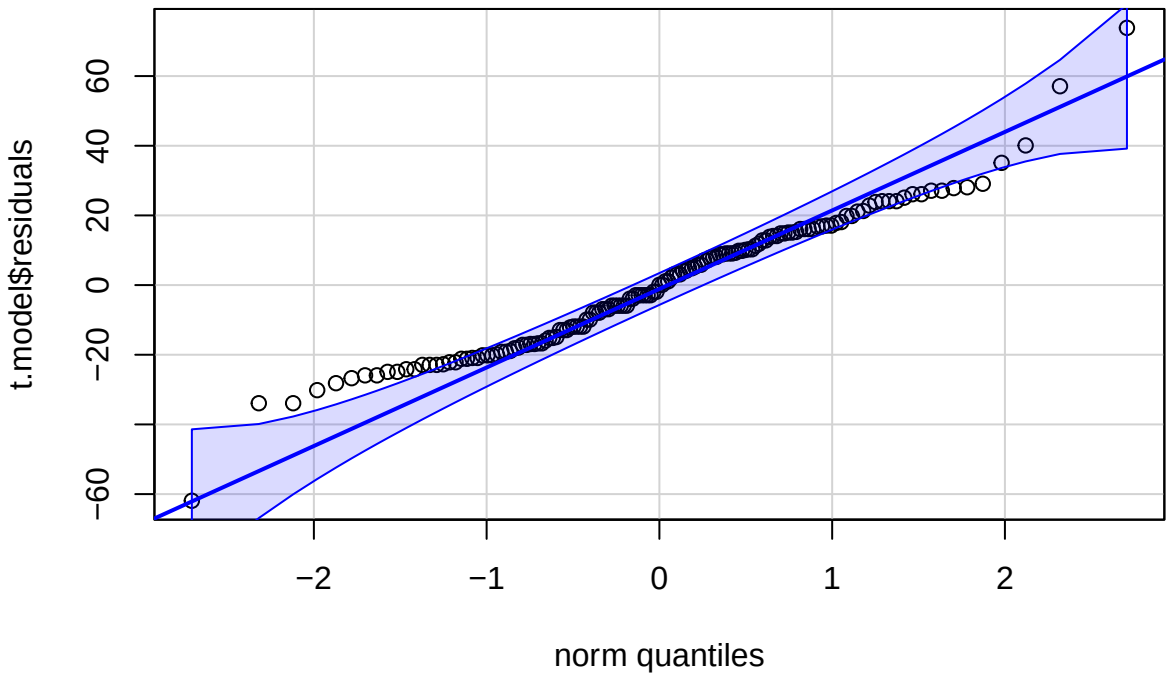
	Df	F value	Pr(>F)
group	2	1.037341	0.3571034
	140	NA	NA

Removing outliers via Studentized Residuals (2)

$\alpha = 0.05$

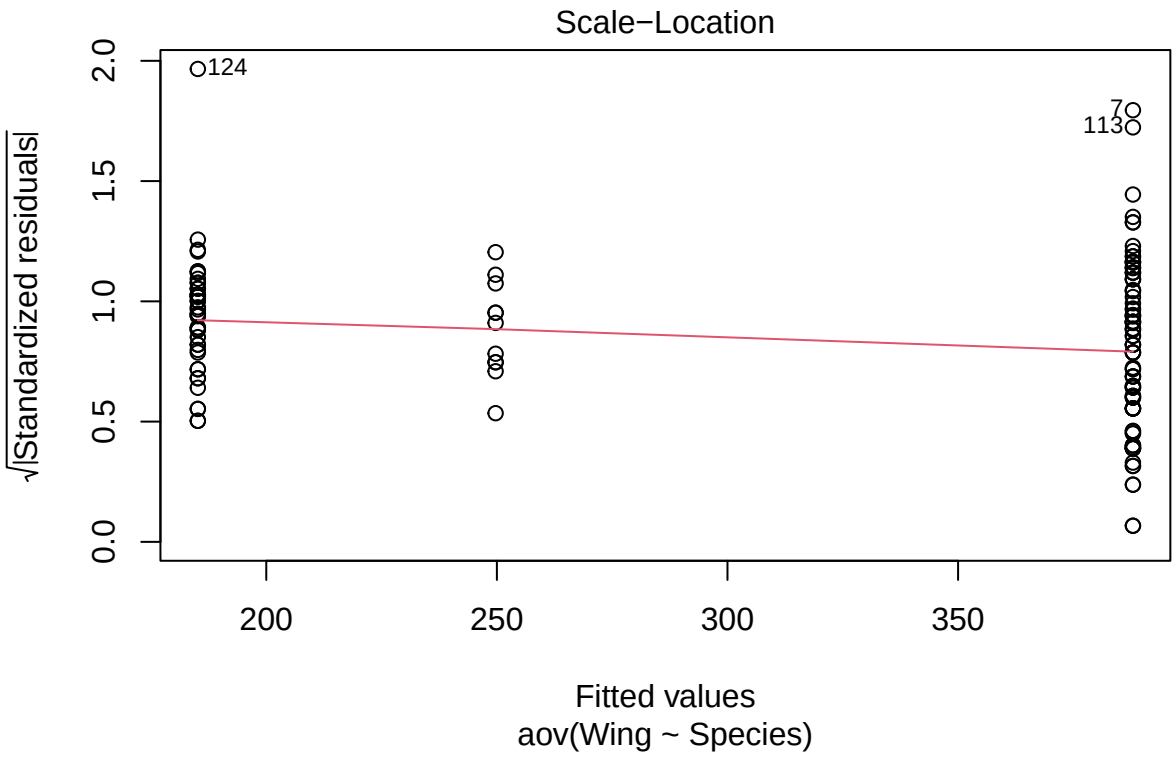
```
##              Df  Sum Sq Mean Sq F value Pr(>F)
## Species       2 1264609  632305    1693 <2e-16 ***
## Residuals   144   53781     373
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Get QQ Plot



Test for normality

Plot variances



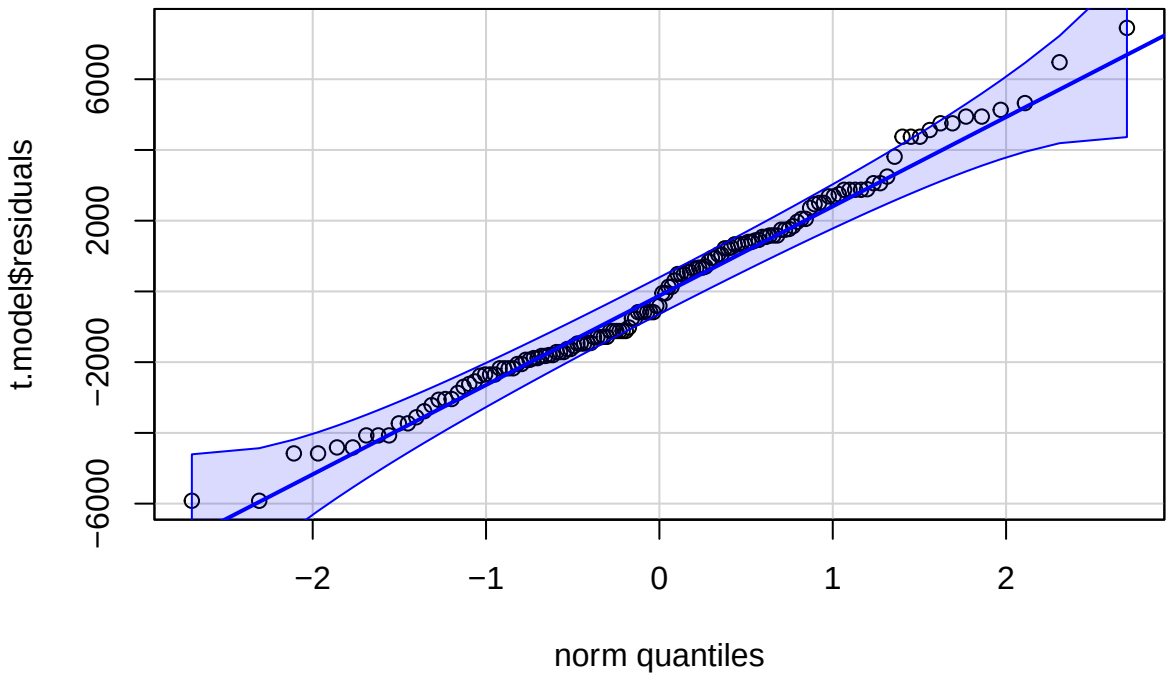
Test for const variance

	Df	F value	Pr(>F)
group	2	0.9823039	0.3769422
	144	NA	NA

Remove outliers (1) and PPCC Transformation

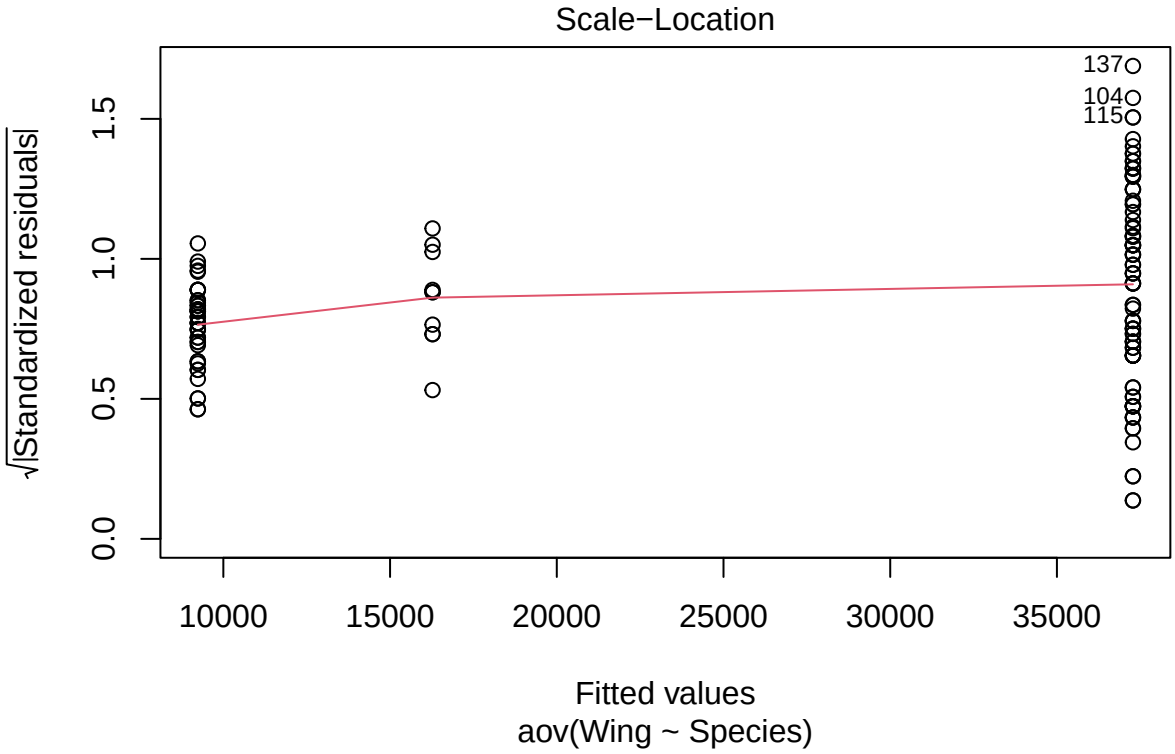
```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## Species       2 2.398e+10 1.199e+10    1737 <2e-16 ***
## Residuals   140 9.660e+08 6.900e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Get QQ Plot



Test for normality

Plot variances



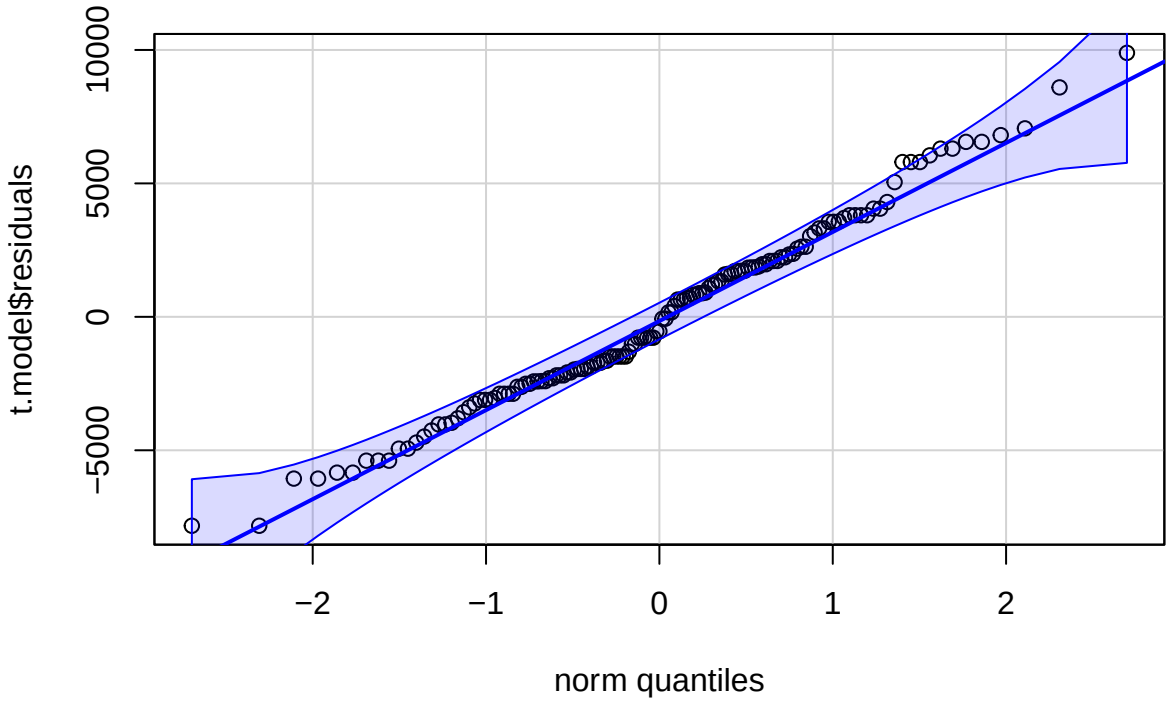
Test for const variance

	Df	F value	Pr(>F)
group	2	4.730775	0.0102778
	140	NA	NA

Remove outliers (1), SW Transformation

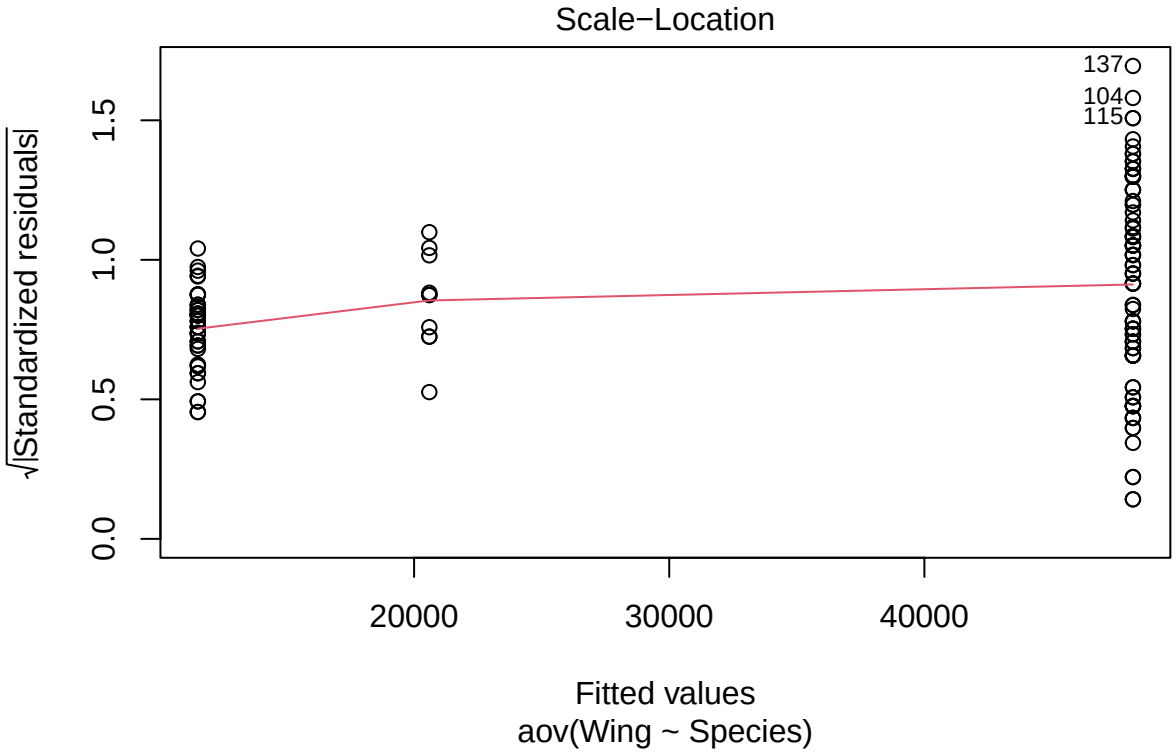
```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## Species       2 4.098e+10 2.049e+10    1709 <2e-16 ***
## Residuals   140 1.679e+09 1.199e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Get QQ Plot



Test for normality

Plot variances



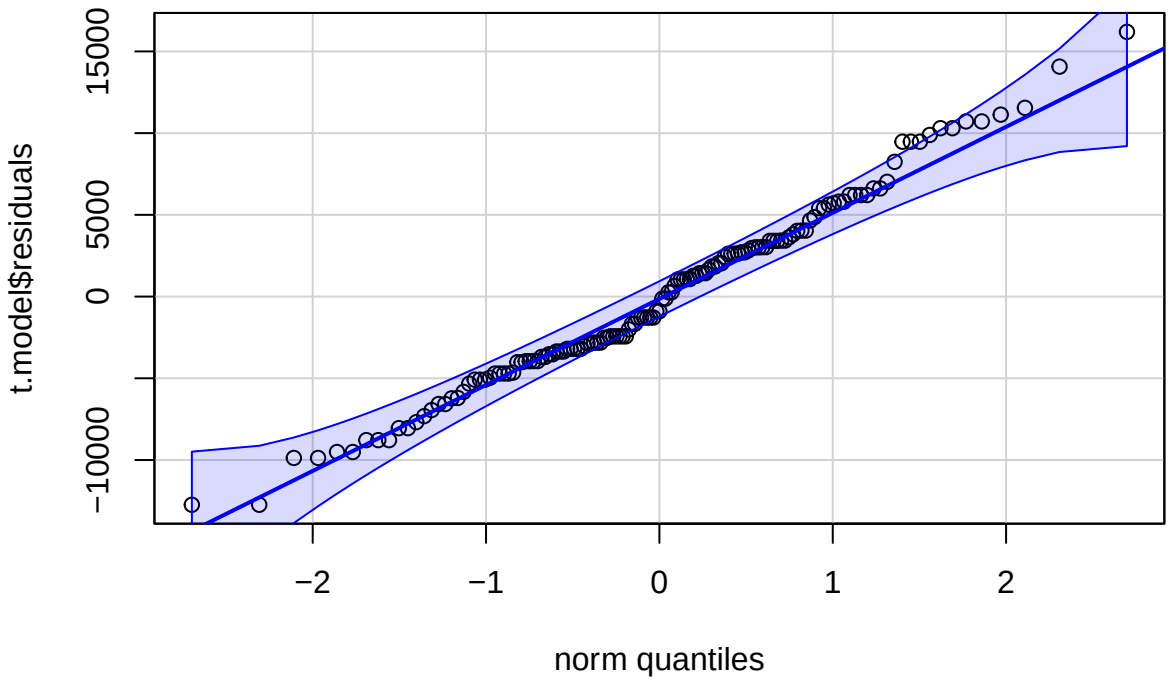
Test for const variance

	Df	F value	Pr(>F)
group	2	5.34138	0.0058145
	140	NA	NA

Remove outliers (1), Log Likelihood Transformation

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## Species       2 1.043e+11 5.215e+10    1658 <2e-16 ***
## Residuals   140 4.402e+09 3.144e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

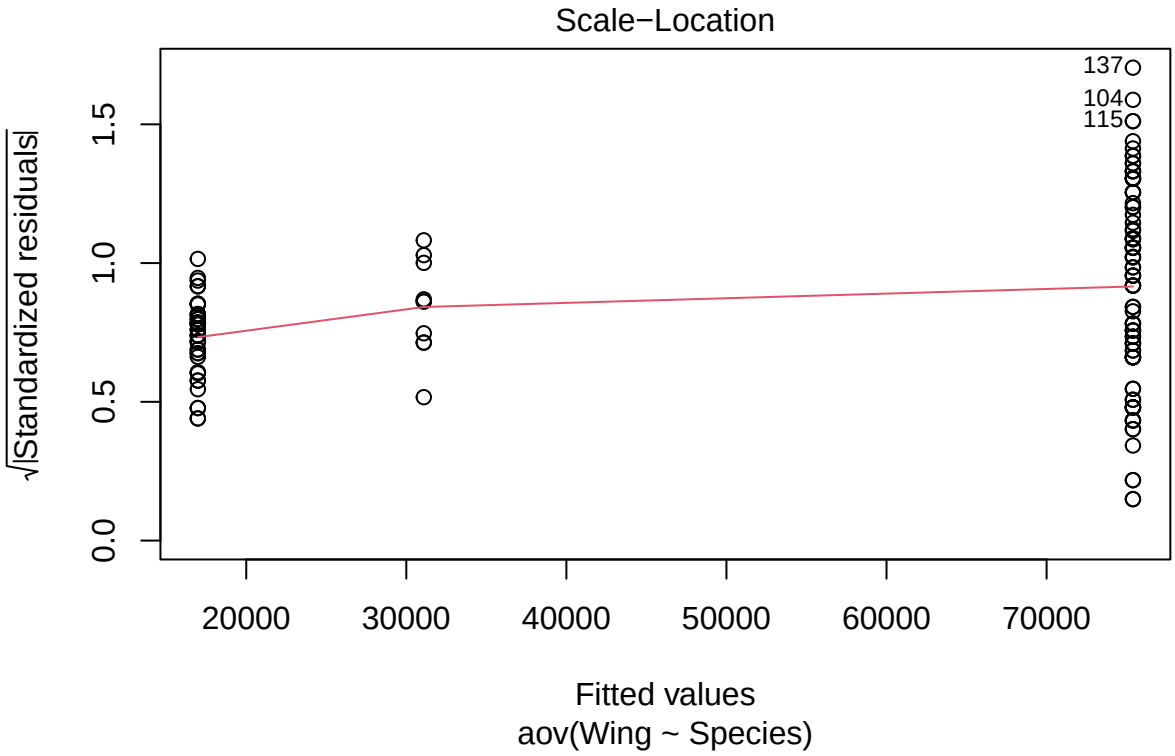
Get QQ Plot



Test for normality

```
##
##  Shapiro-Wilk normality test
##
## data:  t.model$residuals
## W = 0.98669, p-value = 0.1848
```

Plot variances



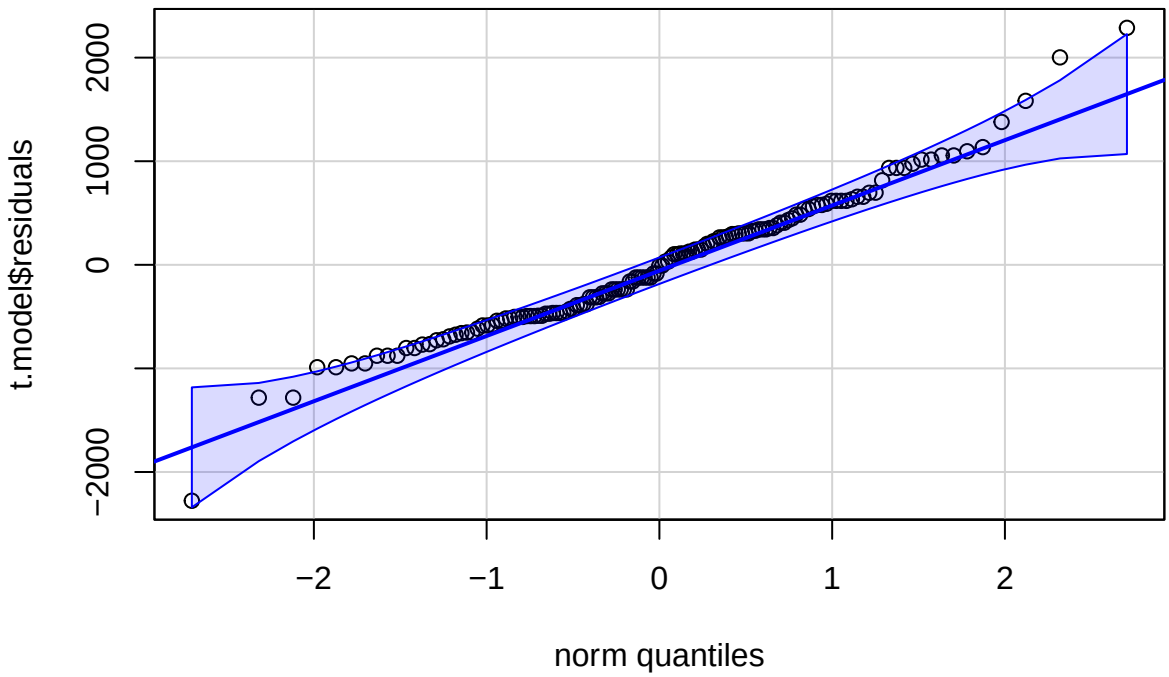
Test for const variance

	Df	F value	Pr(>F)
group	2	6.444978	0.002101
	140	NA	NA

Remove outliers (2) and PPCC Transformation

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## Species       2 1.297e+09 648279640    1504 <2e-16 ***
## Residuals   144 6.207e+07    431030
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

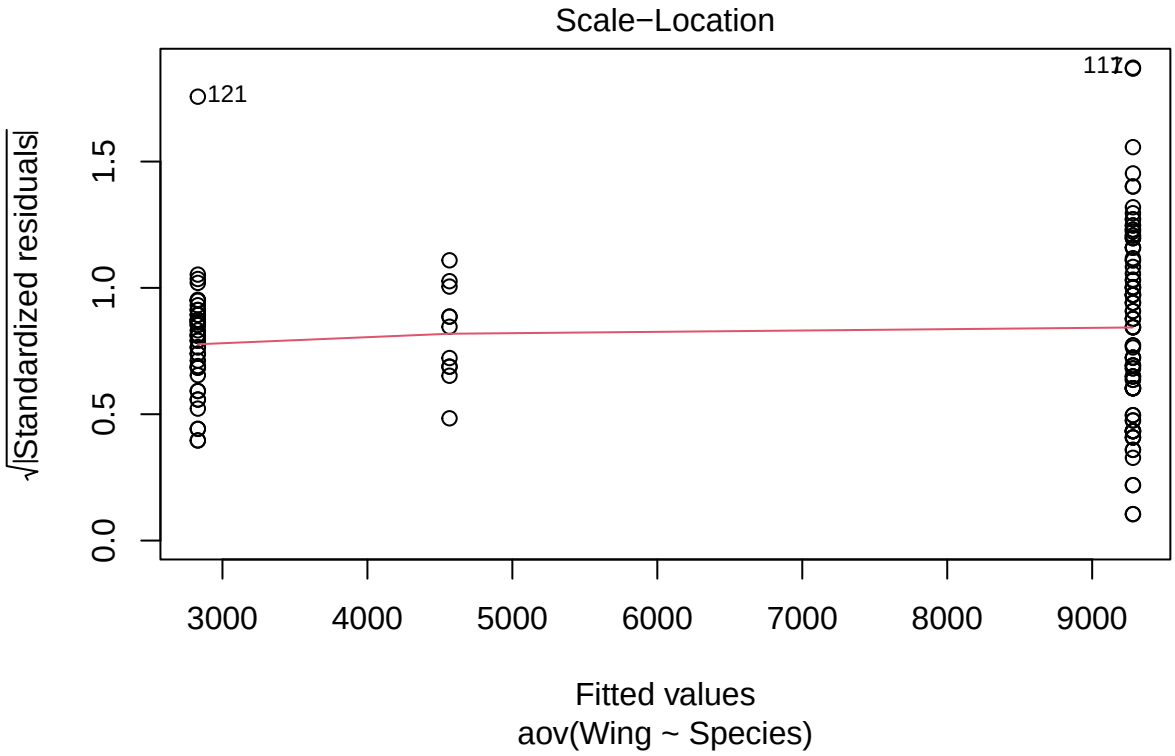
Get QQ Plot



Test for normality

```
##
##  Shapiro-Wilk normality test
##
## data:  t.model$residuals
## W = 0.97616, p-value = 0.01156
```

Plot variances



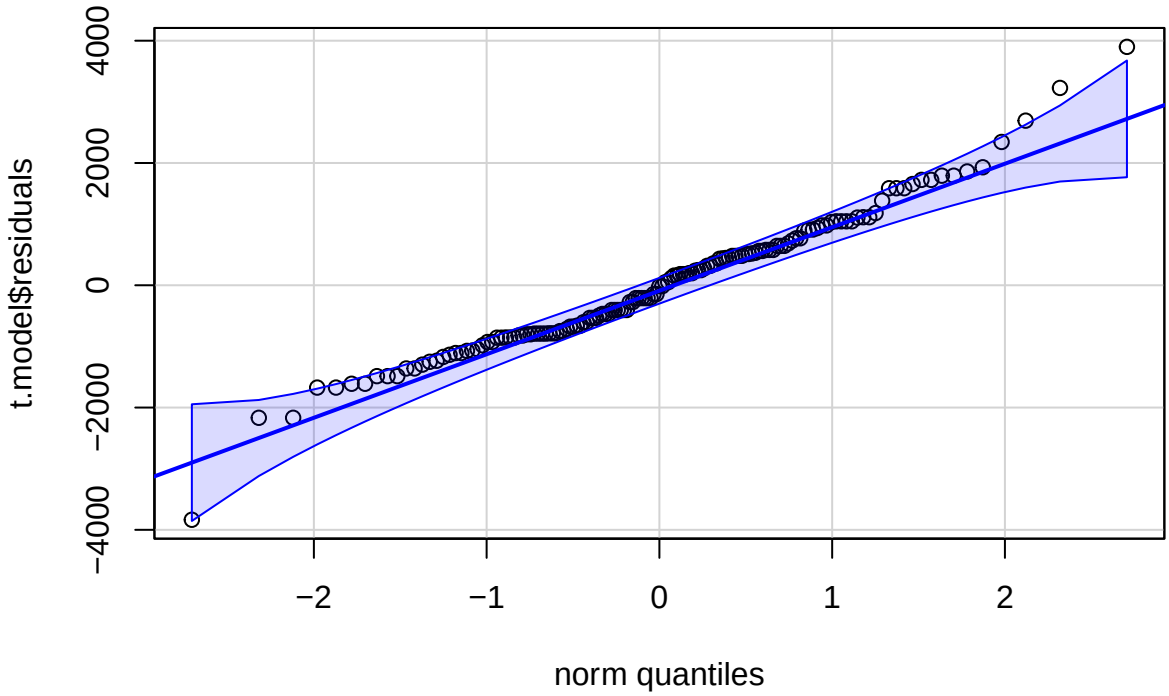
Test for const variance

	Df	F value	Pr(>F)
group	2	1.748162	0.1777679
	144	NA	NA

Remove outliers (2), SW Transformation

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## Species       2 3.542e+09 1.771e+09    1466 <2e-16 ***
## Residuals   144 1.739e+08 1.208e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

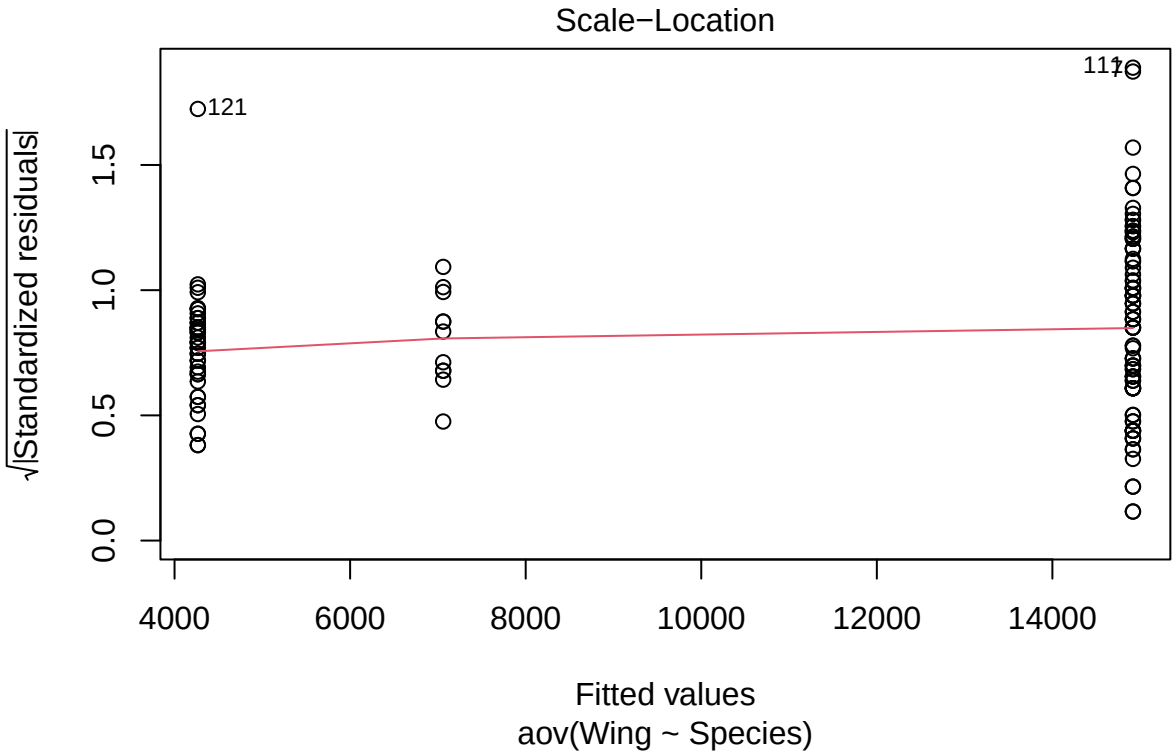
Get QQ Plot



Test for normality

```
##
##  Shapiro-Wilk normality test
##
## data:  t.model$residuals
## W = 0.97624, p-value = 0.01179
```

Plot variances



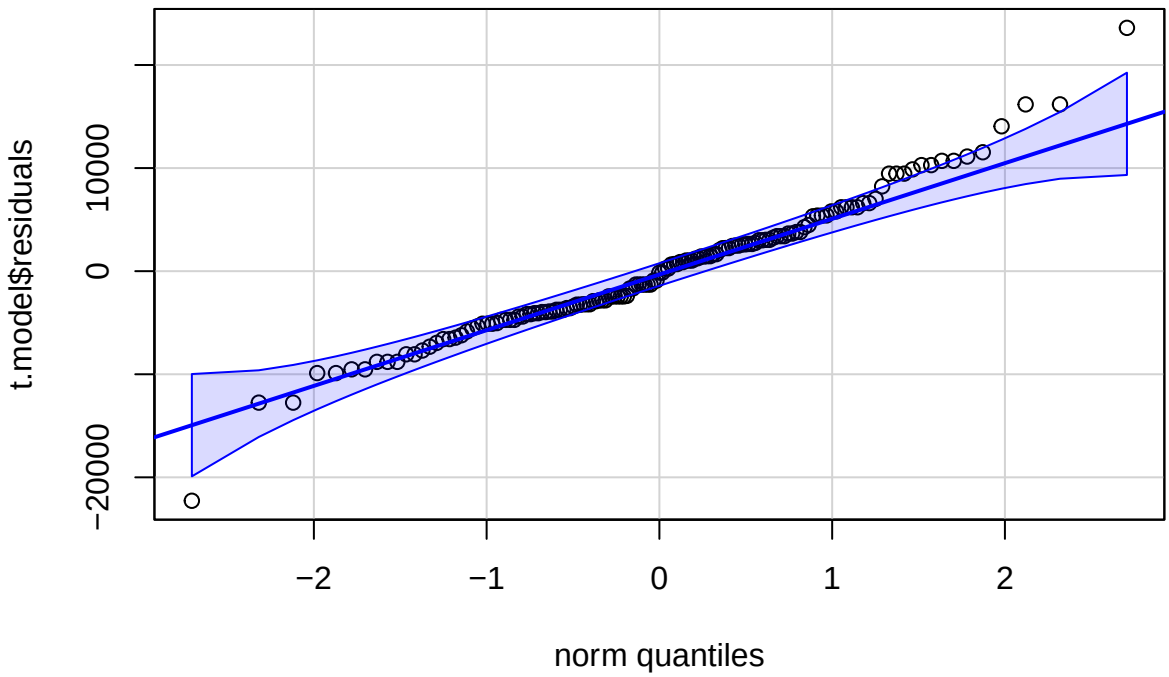
Test for const variance

	Df	F value	Pr(>F)
group	2	2.456738	0.0893
	144	NA	NA

Remove outliers (2), Log Likelihood Transformation

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## Species       2 1.058e+11 5.291e+10    1330 <2e-16 ***
## Residuals   144 5.729e+09 3.978e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

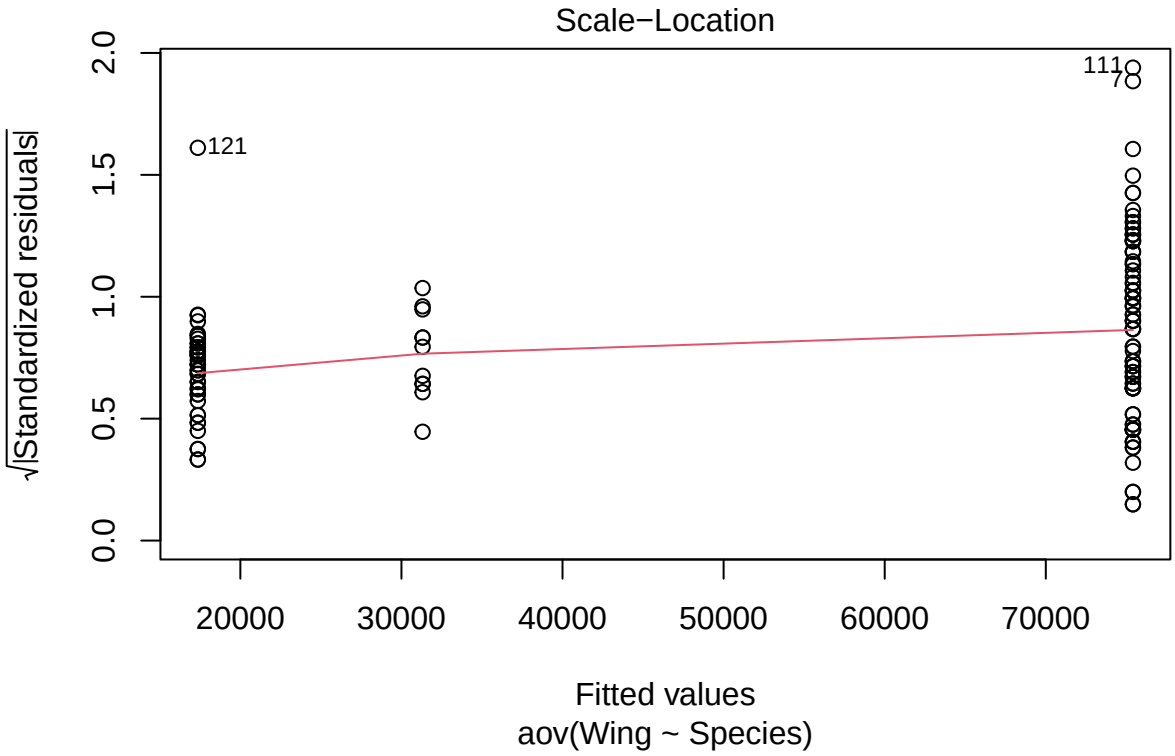
Get QQ Plot



Test for normality

```
##
##  Shapiro-Wilk normality test
##
## data:  t.model$residuals
## W = 0.974, p-value = 0.0068
```

Plot variances



Test for const variance

	Df	F value	Pr(>F)
group	2	5.453402	0.0052122
	144	NA	NA

Conclusion

There are 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.

Appendix

```r
# Import dataset
the.data <- read.csv("NewHawk.csv")
the.model <- lm(Wing ~ Species, data = the.data)
model.fit <- aov(Wing ~ Species,
  data = the.data
)

# Set size of all plots
options(
  repr.plot.width = 4,  # Width of the plot in inches
  repr.plot.height = 3  # Height of the plot in inches
)
```

```r
library("ggplot2")
ggplot(the.data, aes(x = Wing, fill = Species)) +
  geom_histogram(binwidth = 10, color = "black", fill = "white") +
  facet_grid(Species ~ .) +
  labs(title = "Figure 1.1.1 Histogram of Wing Feather Length by Group")
```

```r
ei = model.fit$residuals
boxplot(ei ~ Species, data = the.data,
        main = "Figure 1.1.2 Box Plot of Different Species")
```

```r
car::qqPlot(model.fit$residuals,
  id = FALSE,
  main = "Figure 1.2 QQ Plot of Original Data"
)
```

```r
the.SWtest = shapiro.test(model.fit$residuals)
p_value <- round(the.SWtest$p.value, 4)
```

```r
plot(model.fit, which = 3, main = "Figure 1.3 Error Plot of Original Data")
```

```r
levene_test <- car::leveneTest(model.fit)
p_value <- round(levene_test[1, 3], 4)
```

```r
# All transformations considered
#QQplot
L1 <- EnvStats::boxcox(model.fit, objective.name = "PPCC", optimize = TRUE)$lambda
#Shapiro-Wilks
L2 <- EnvStats::boxcox(model.fit, objective.name = "Shapiro-Wilk", optimize = TRUE)$lambda
L3 <- EnvStats::boxcox(the.data$Wing, objective.name = "Log-Likelihood",optimize = TRUE)$lambda
```

```r
par(mfrow = c(1, 2))

# The transformation function to get transformed model
give.me.t.model <- function(L, data = the.data) {
  YT = (data$Wing^(L) - 1) / L
```

```
  t.data = data.frame(Wing = YT, Species = data$Species)
  t.model = aov(Wing ~ Species, data = t.data)
  return(t.model)
}
```

```
# Get summary of this model
t.model <- give.me.t.model(L1)
# summary(t.model)
```

```
car::qqPlot(t.model$residuals,
  id = FALSE,
  main = "QQPlot"
)
```

```
the.SWtest = shapiro.test(t.model$residuals)
the.p.val <- round(the.SWtest$p.value, 4)
```

```
# Homogeneity of variances
plot(t.model, which = 3)
```

```
# Levene test
levene_test = car::leveneTest(t.model)
p_value <- round(levene_test[1, 3], 4)
```

```
# Get summary of this model
t.model <- give.me.t.model(L3)
summary(t.model)
```

```
car::qqPlot(t.model$residuals,
id = FALSE # remove point identification
)
```

```
the.SWtest = shapiro.test(t.model$residuals)
the.p.val <- round(the.SWtest$p.value, 4)
```

```
# Homogeneity of variances
plot(t.model, which = 3)
```

```
# Levene test
car::leveneTest(t.model)
```

```
# Get summary of this model
t.model <- give.me.t.model(L2)
summary(t.model)
```

```
car::qqPlot(t.model$residuals,
id = FALSE # remove point identification
)
```

```r
the.SWtest = shapiro.test(t.model$residuals)
the.p.val <- round(the.SWtest$p.value, 4)
```

```r
# Homogeneity of variances
plot(t.model, which = 3)
```

```r
# Levene test
car::leveneTest(t.model)
```

```r
# remove outliers
the.box.plot <- boxplot(Wing ~ Species, data = the.data)
outliers <- the.box.plot$out
data.no.outlier.1 <- the.data[!the.data$Wing %in% outliers, ]
```

```r
# Get summary of this model
t.model <- aov(Wing ~ Species, data = data.no.outlier.1)
model.no.outlier.1 <- t.model
summary(t.model)
```

```r
car::qqPlot(t.model$residuals,
  id = FALSE # remove point identification
)
```

```r
the.SWtest = shapiro.test(t.model$residuals)
the.SWtest
```

```r
# Homogeneity of variances
plot(t.model, which = 3)
```

```r
# Levene test
car::leveneTest(t.model)
```

```r
rij = rstandard(the.model)
alpha = 0.05

nt = nrow(the.data) #Calculates the total sample size
a = length(unique(the.data$Species)) #Calculates the value of a
t.cutoff= qt(1-alpha/(2*nt), nt-a)
CO.rij = which(abs(rij) > t.cutoff)

outliers = CO.rij
data.no.outlier.2 = the.data[-outliers,]
```

```r
# Get summary of this model
t.model <- aov(Wing ~ Species, data = data.no.outlier.2)
model.no.outlier.2 <- t.model
summary(t.model)
```

```r
car::qqPlot(t.model$residuals,
  id = FALSE # remove point identification
)
```

```r
the.SWtest = shapiro.test(t.model$residuals)
the.p.val <- round(the.SWtest$p.value, 4)
```

```r
# Homogeneity of variances
plot(t.model, which = 3)
```

```r
# Levene test
car::leveneTest(t.model)
```

```r
# All transformations considered
#QQplot
L1 <- EnvStats::boxcox(model.no.outlier.1, objective.name = "PPCC", optimize = TRUE)$lambda
#Shapiro-Wilks
L2 <- EnvStats::boxcox(model.no.outlier.1, objective.name = "Shapiro-Wilk", optimize = TRUE)$lambda
L3 <- EnvStats::boxcox(data.no.outlier.1$Wing, objective.name = "Log-Likelihood",optimize = TRUE)$lambda
```

```r
# Get summary of this model
t.model <- give.me.t.model(L1, data = data.no.outlier.1)
summary(t.model)
```

```r
car::qqPlot(t.model$residuals,
id = FALSE # remove point identification
)
```

```r
the.SWtest = shapiro.test(t.model$residuals)
the.p.val <- round(the.SWtest$p.value, 4)
```

```r
# Homogeneity of variances
plot(t.model, which = 3)
```

```r
# Levene test
car::leveneTest(t.model)
```

```r
# Get summary of this model
t.model <- give.me.t.model(L2, data = data.no.outlier.1)
summary(t.model)
```

```r
car::qqPlot(t.model$residuals,
id = FALSE # remove point identification
)
```

```r
the.SWtest = shapiro.test(t.model$residuals)
the.p.val <- round(the.SWtest$p.value, 4)
```

```r
# Homogeneity of variances
plot(t.model, which = 3)
```

```r
# Levene test
car::leveneTest(t.model)
```

```r
# Get summary of this model
t.model <- give.me.t.model(L3, data = data.no.outlier.1)
summary(t.model)
```

```r
car::qqPlot(t.model$residuals,
id = FALSE # remove point identification
)
```

```r
the.SWtest = shapiro.test(t.model$residuals)
the.SWtest
```

```r
# Homogeneity of variances
plot(t.model, which = 3)
```

```r
# Levene test
car::leveneTest(t.model)
```

```r
# All transformations considered
#QQplot
L1 <- EnvStats::boxcox(model.no.outlier.2, objective.name = "PPCC", optimize = TRUE)$lambda
#Shapiro-Wilks
L2 <- EnvStats::boxcox(model.no.outlier.2, objective.name = "Shapiro-Wilk", optimize = TRUE)$lambda
L3 <- EnvStats::boxcox(data.no.outlier.2$Wing, objective.name = "Log-Likelihood",optimize = TRUE)$lambda
```

```r
# Get summary of this model
t.model <- give.me.t.model(L1, data = data.no.outlier.2)
summary(t.model)
```

```r
car::qqPlot(t.model$residuals,
id = FALSE # remove point identification
)
```

```r
the.SWtest = shapiro.test(t.model$residuals)
the.SWtest
```

```r
# Homogeneity of variances
plot(t.model, which = 3)
```

```r
# Levene test
car::leveneTest(t.model)
```

```r
# Get summary of this model
t.model <- give.me.t.model(L2, data = data.no.outlier.2)
summary(t.model)
```

```r
car::qqPlot(t.model$residuals,
id = FALSE # remove point identification
)
```

```r
the.SWtest = shapiro.test(t.model$residuals)
the.SWtest
```

```r
# Homogeneity of variances
plot(t.model, which = 3)
```

```r
# Levene test
car::leveneTest(t.model)
```

```r
# Get summary of this model
t.model <- give.me.t.model(L3, data = data.no.outlier.2)
summary(t.model)
```

```r
car::qqPlot(t.model$residuals,
id = FALSE # remove point identification
)
```

```r
the.SWtest = shapiro.test(t.model$residuals)
the.SWtest
```

```r
# Homogeneity of variances
plot(t.model, which = 3)
```

```r
# Levene test
car::leveneTest(t.model)
```