Project 2 Outliers and Transformation

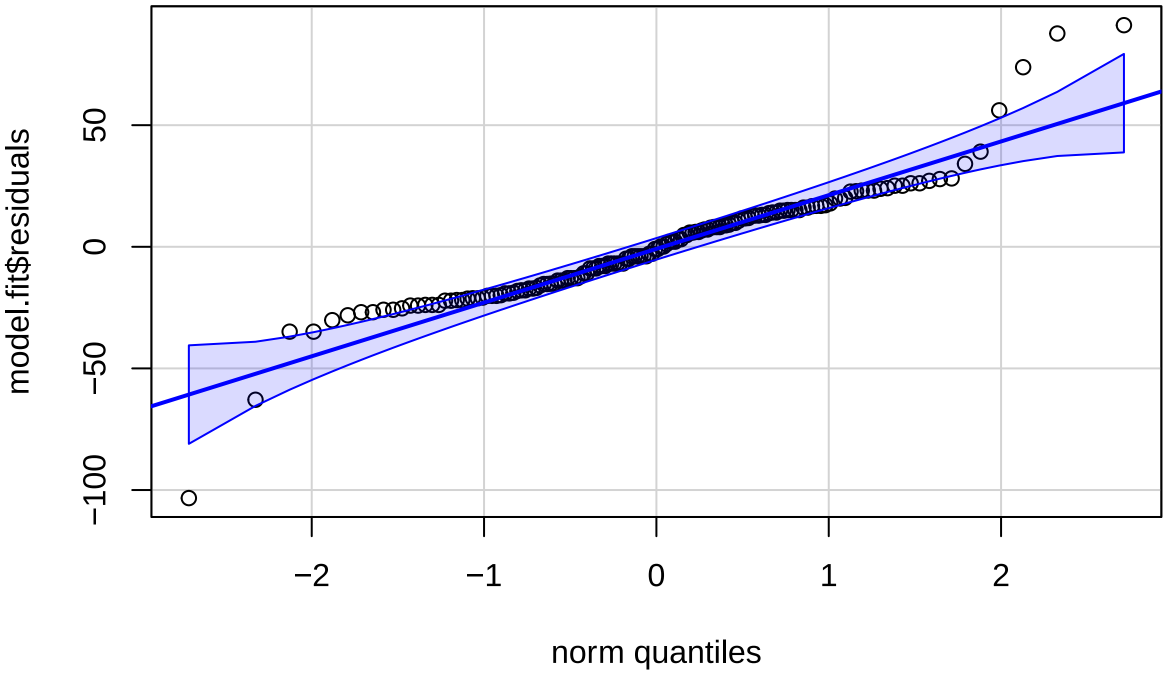
Andrew Jowe

Assumption

$\alpha = 0.1$ for both const variance and normality test.

Import dataset and get model fit

Get QQ Plot

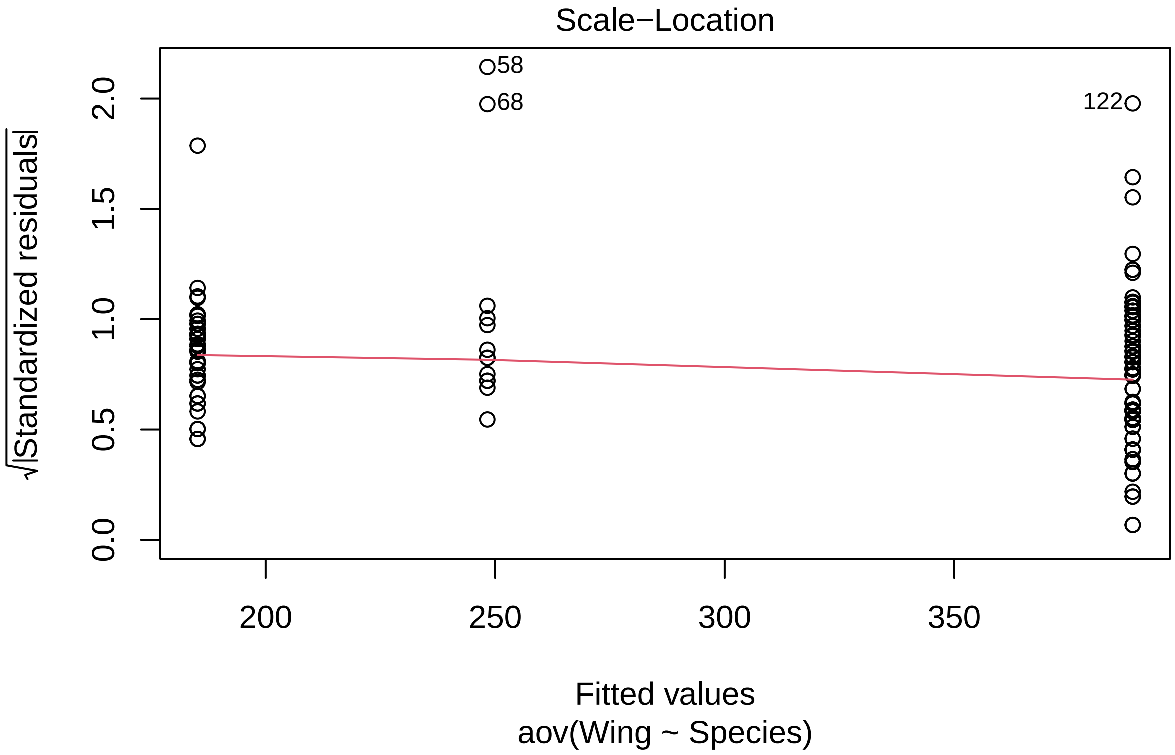


Do test for normality

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  model.fit$residuals
## W = 0.91732, p-value = 1.431e-07
```

Do plot for variance



Do test for const variance

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

	Df	F value	Pr(>F)
group	2	2.348052	0.0991297
	147	NA	NA

Transformation Possibilities

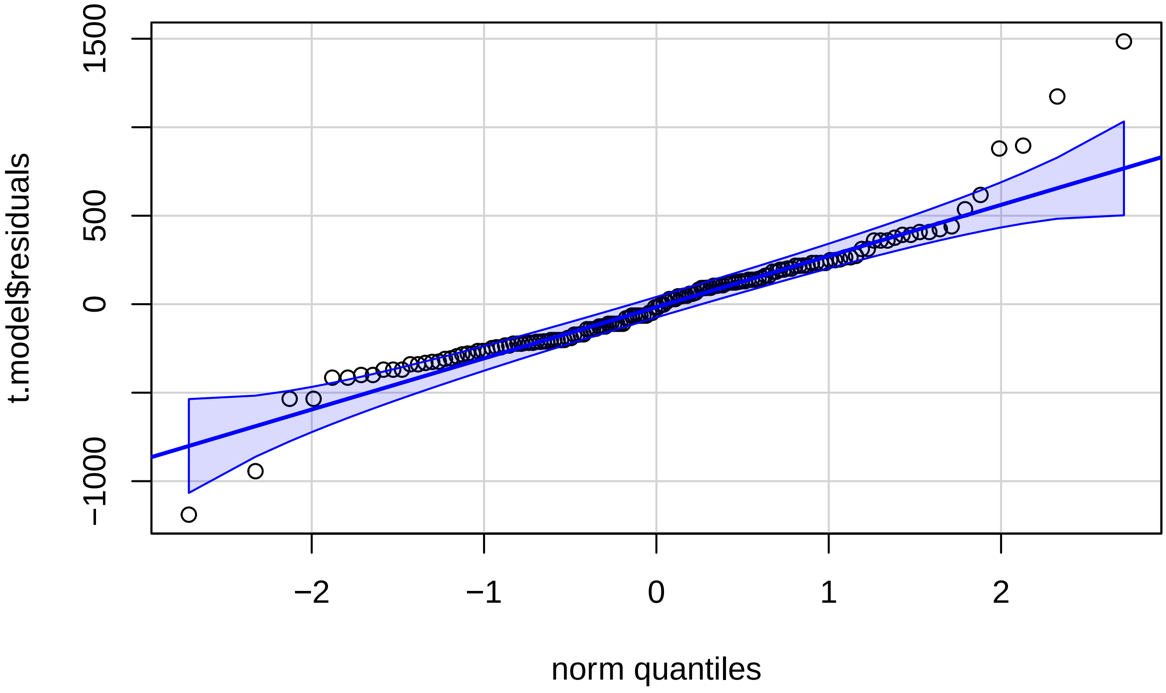
1. PPCC

2. Shapiro-Wilks
3. Log-Likelihood
4.

No outlier removal, PPCC Transformation

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## Species         2 235559223 117779611    1111 <2e-16 ***
## Residuals     147  15589809    106053
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

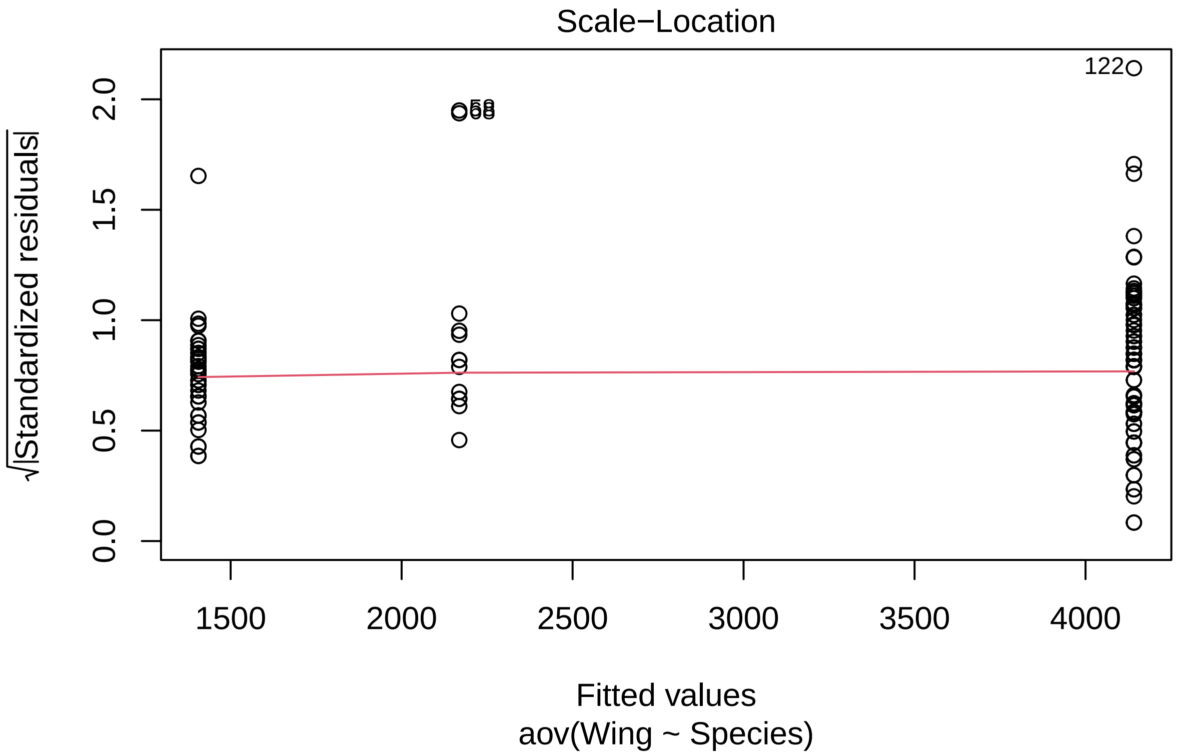
Get QQ Plot



Test for normality

```
##
##  Shapiro-Wilk normality test
##
## data:  t.model$residuals
## W = 0.92327, p-value = 3.473e-07
```

Plot variances



Test for const variance

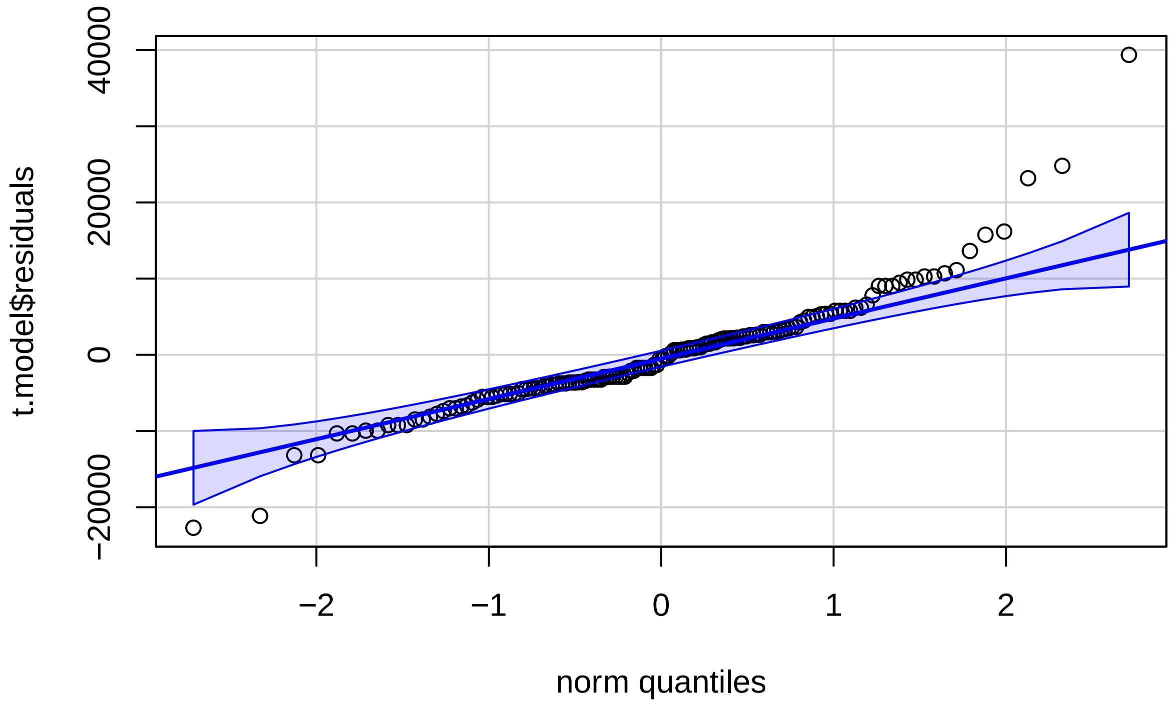
```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

	Df	F value	Pr(>F)
group	2	1.626697	0.2000961
	147	NA	NA

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
```

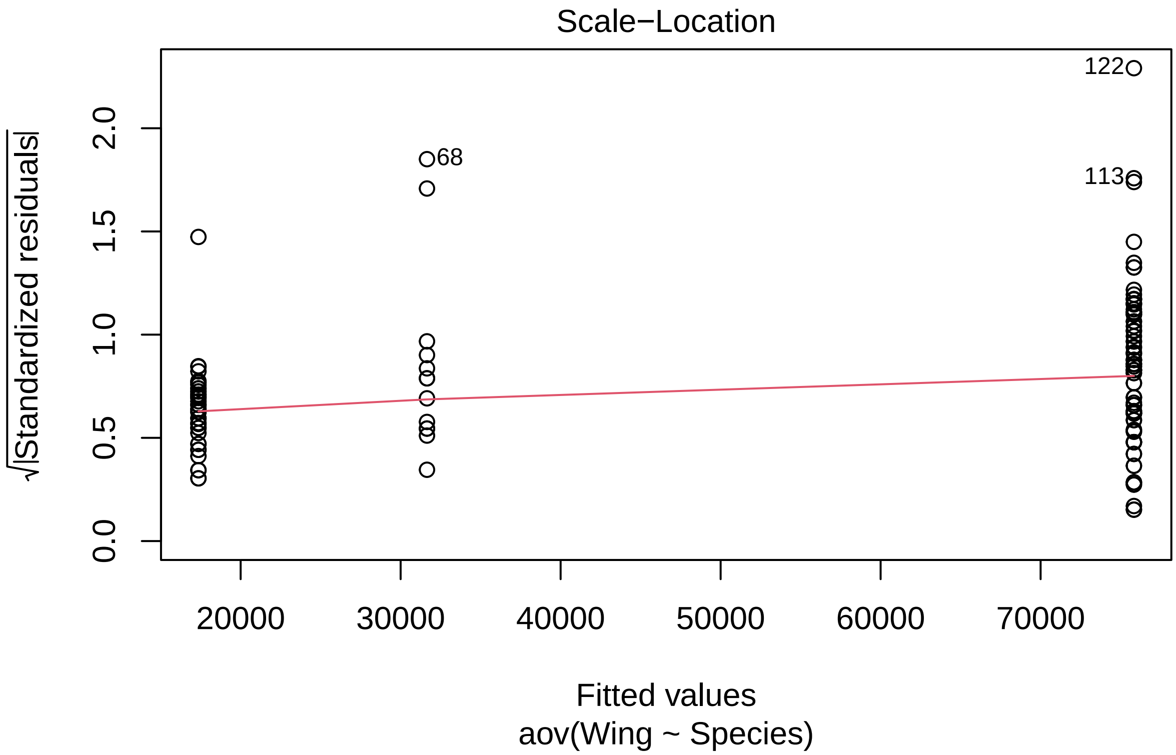
Get QQ Plot



Test for normality

```
##
##  Shapiro-Wilk normality test
##
## data:  t.model$residuals
## W = 0.91351, p-value = 8.275e-08
```

Plot variances



Test for const variance

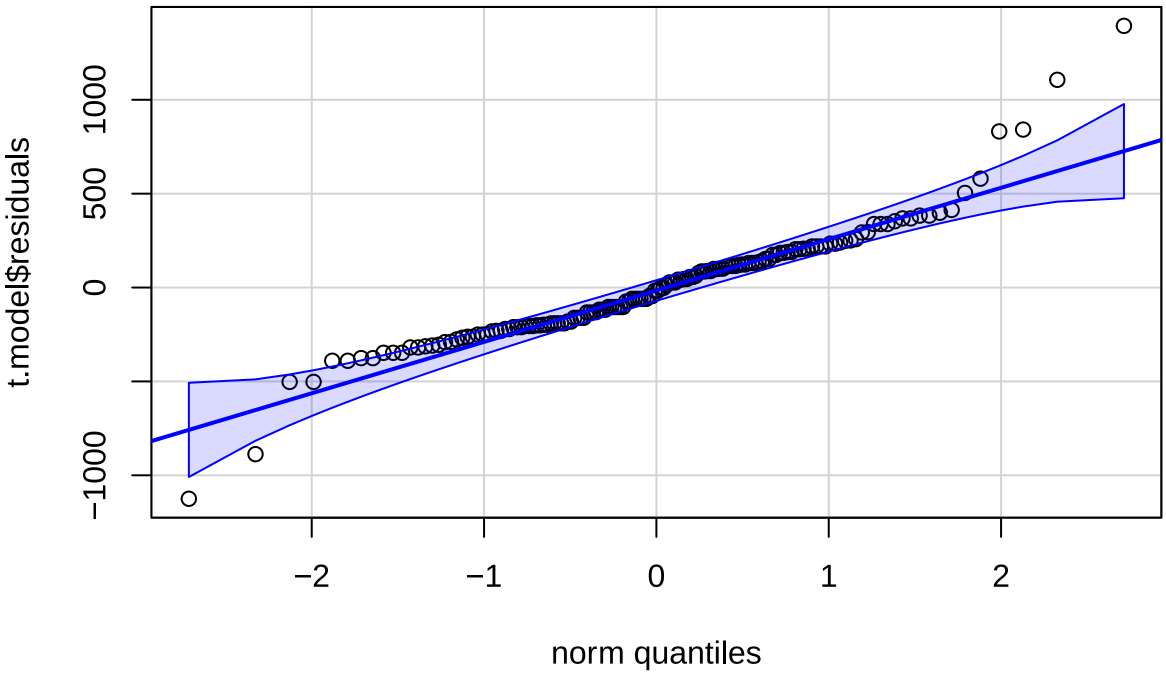
```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

	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 209280777 104640388    1113 <2e-16 ***
## Residuals   147  13820960     94020                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

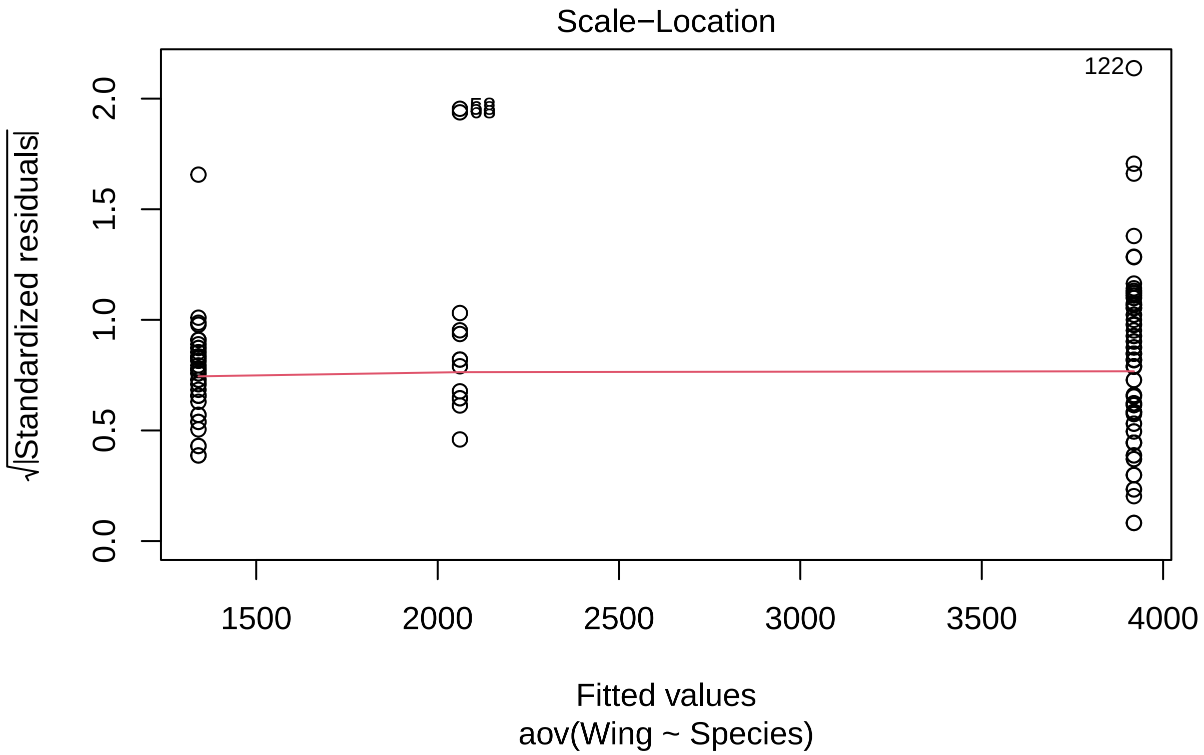
Get QQ Plot



Test for normality

```
## 
##  Shapiro-Wilk normality test
## 
## data:  t.model$residuals
## W = 0.92328, p-value = 3.476e-07
```

Plot variances



Test for const variance

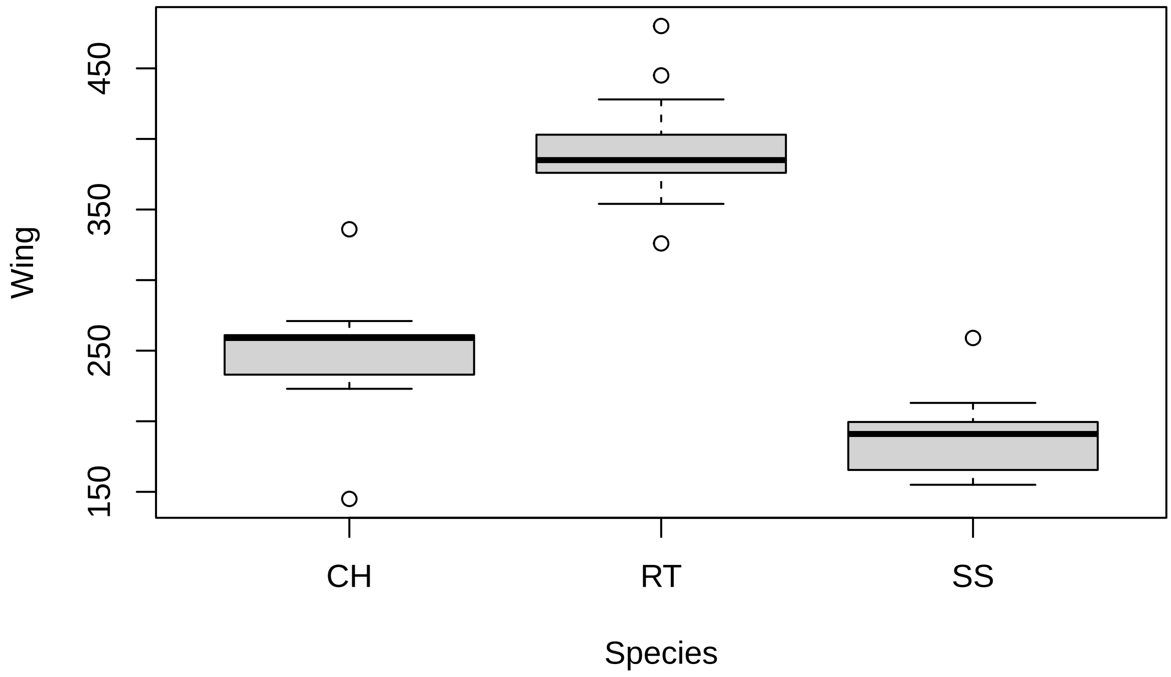
```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

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

Possible outlier removal techniques

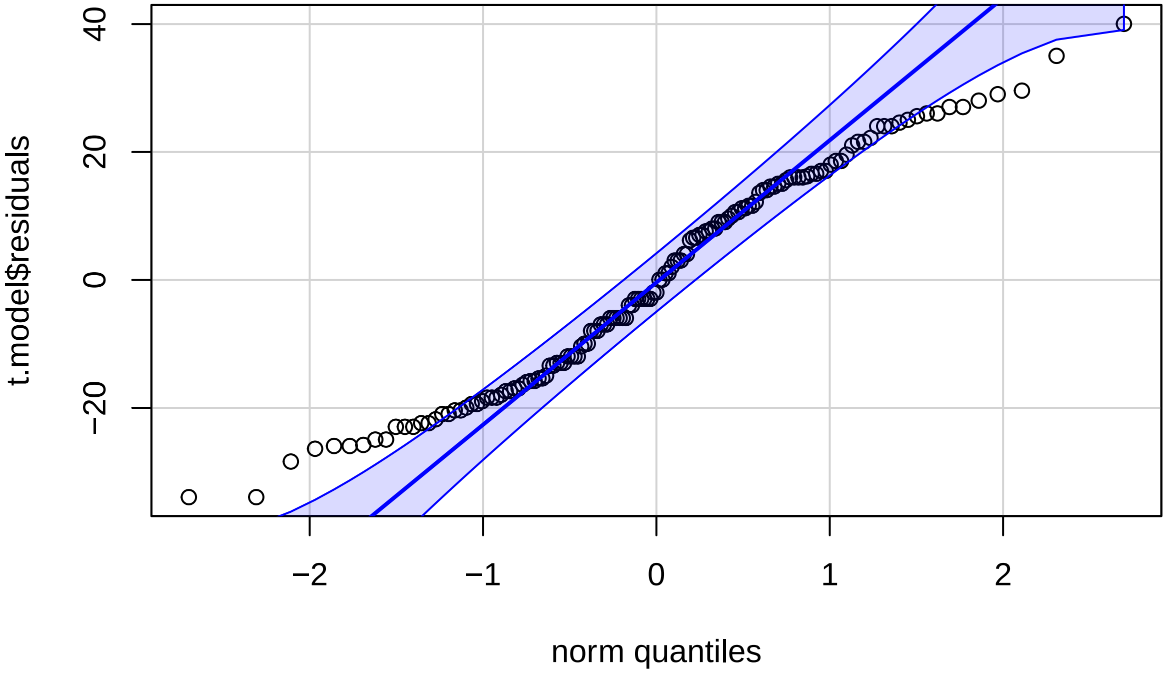
1. Outlier removal via box plot
2. Semi-Studentized Residuals: we can use this since we have the assumption that our variance is constant from our original test. We don't need to do studentized residuals since this is a more robust replacement.

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
```

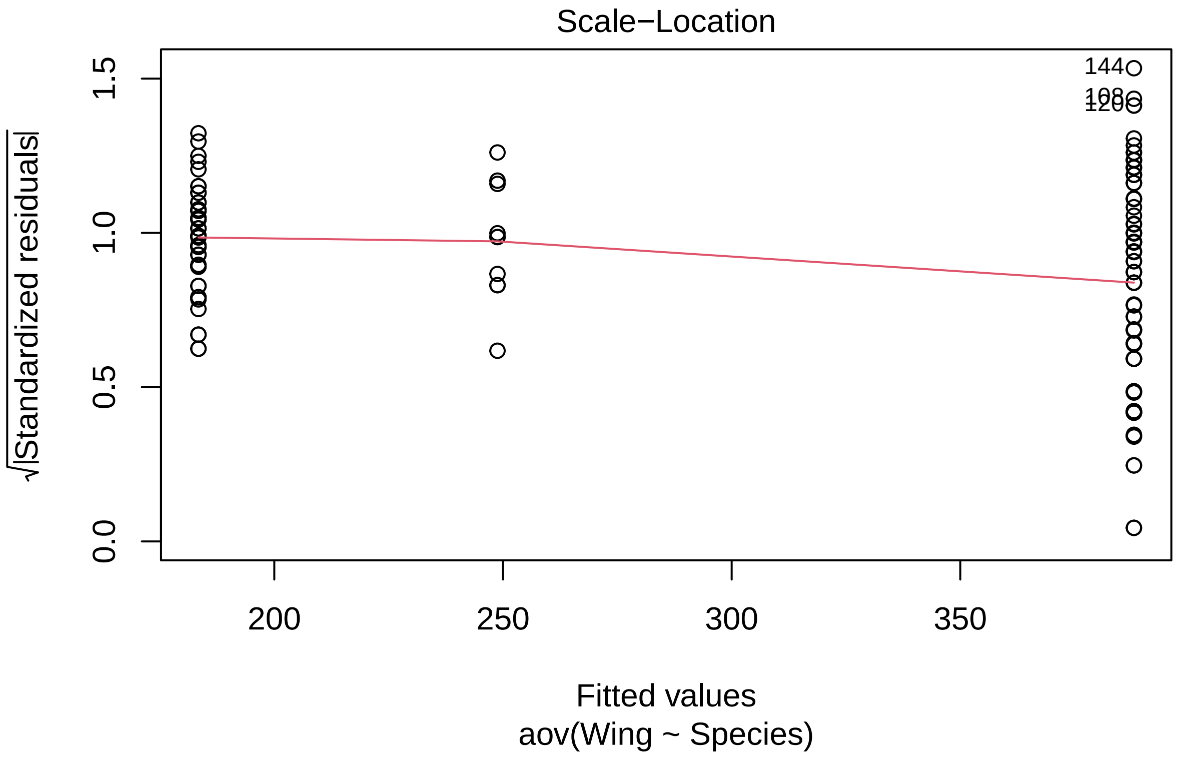
Get QQ Plot



Test for normality

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

Plot variances



Test for const variance

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

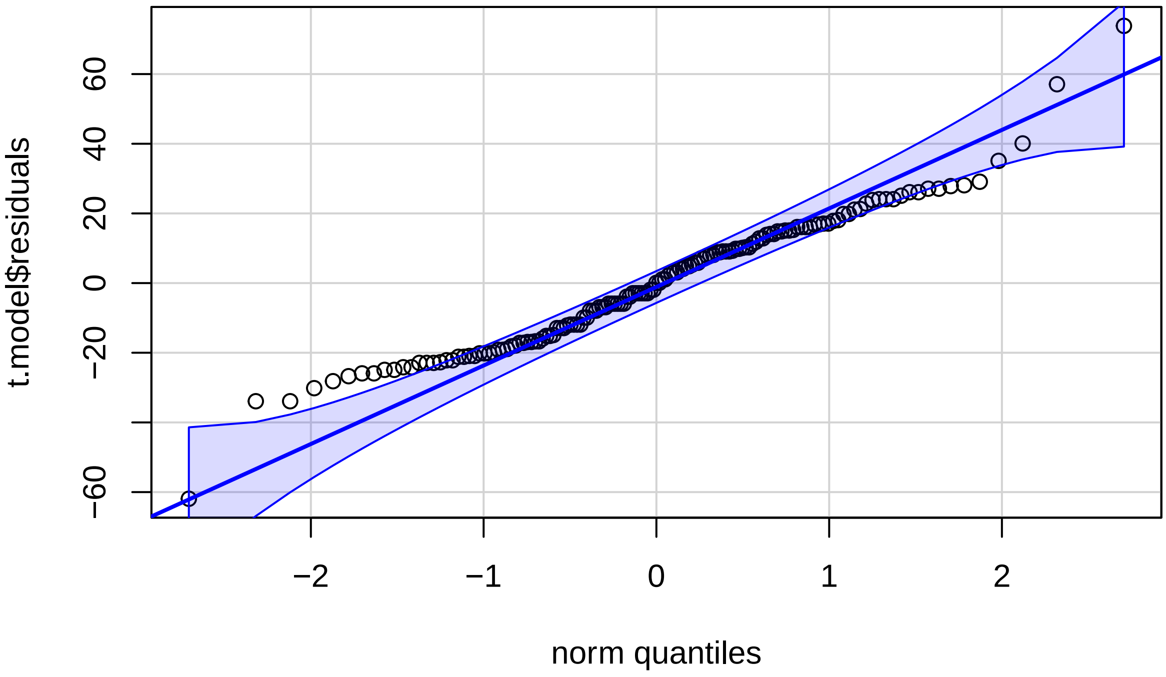
	Df	F value	Pr(>F)
group	2	1.037341	0.3571034
	140	NA	NA

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
```

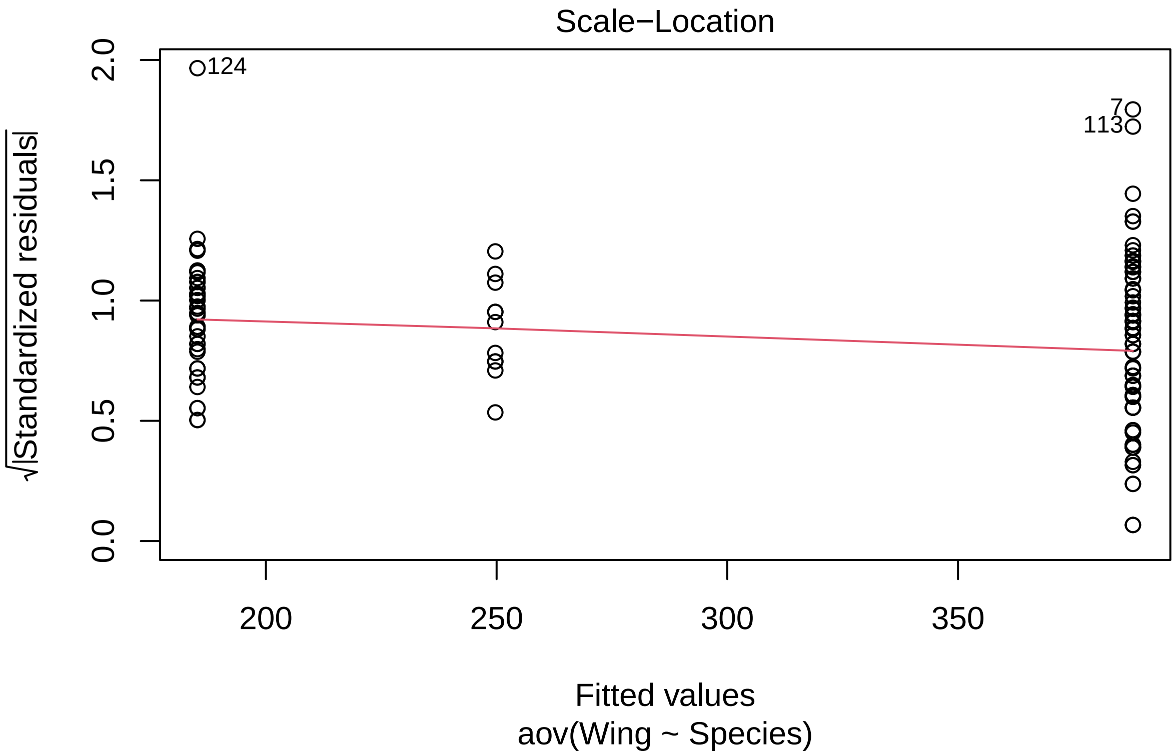
Get QQ Plot



Test for normality

```
##
##  Shapiro-Wilk normality test
##
## data:  t.model$residuals
## W = 0.97182, p-value = 0.004021
```

Plot variances



Test for const variance

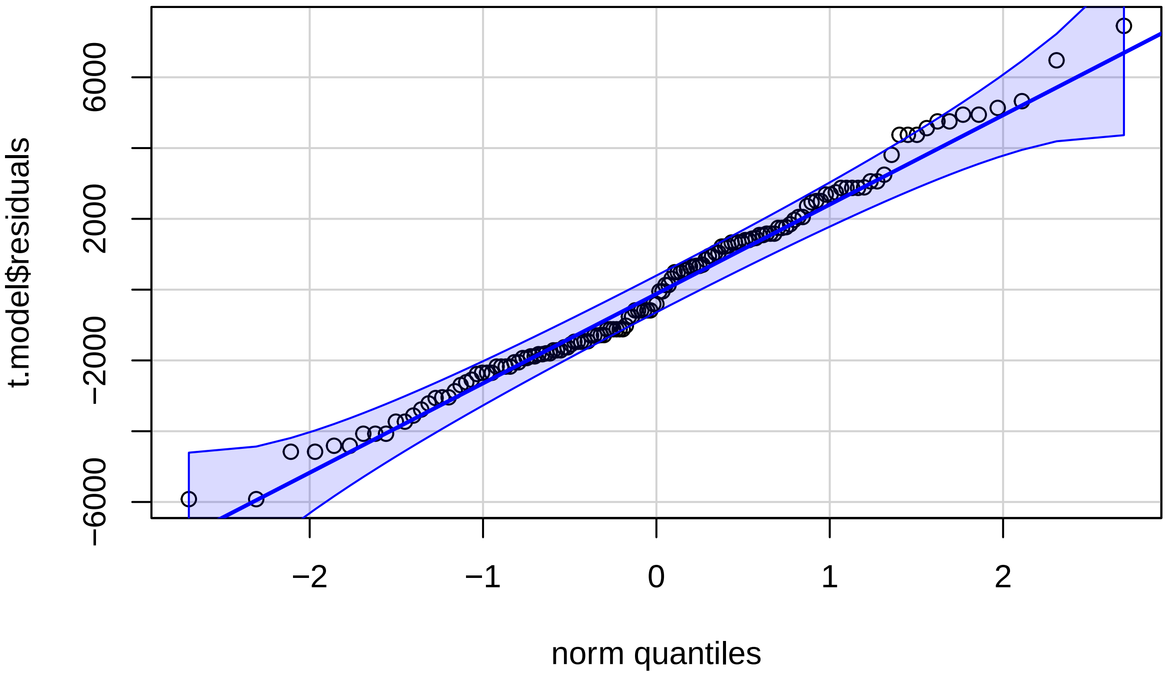
```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

	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
```

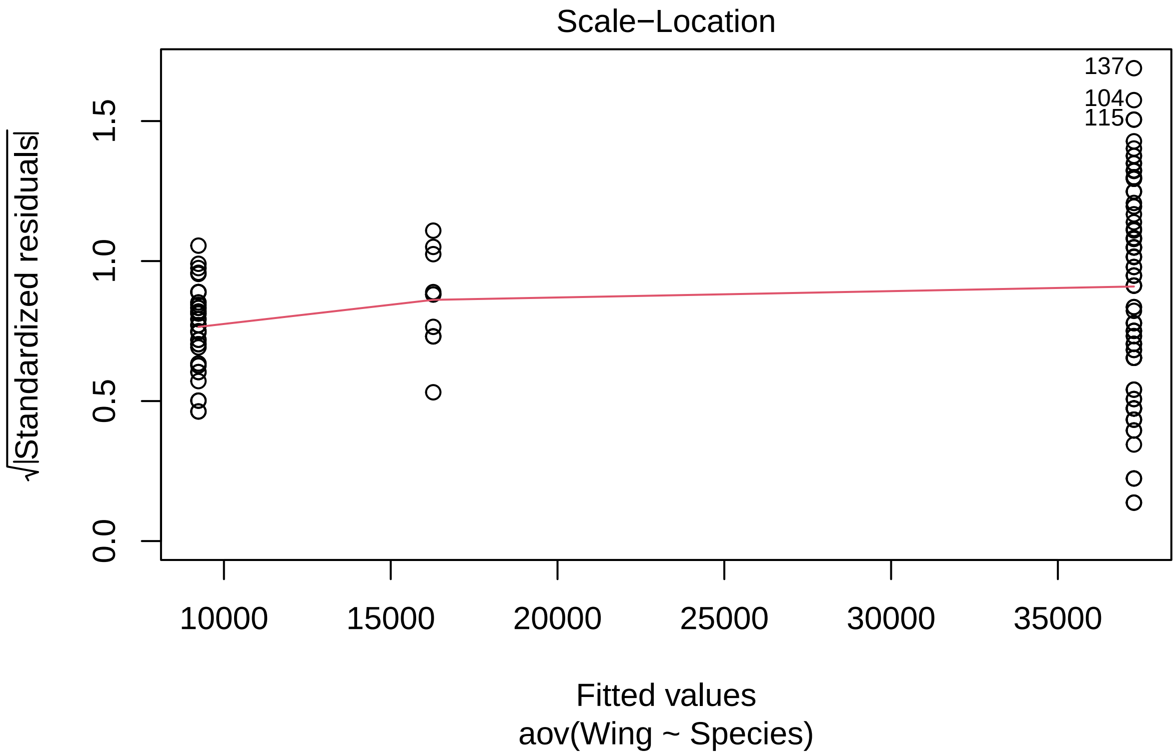
Get QQ Plot



Test for normality

```
##
##  Shapiro-Wilk normality test
##
## data:  t.model$residuals
## W = 0.98691, p-value = 0.1952
```

Plot variances



Test for const variance

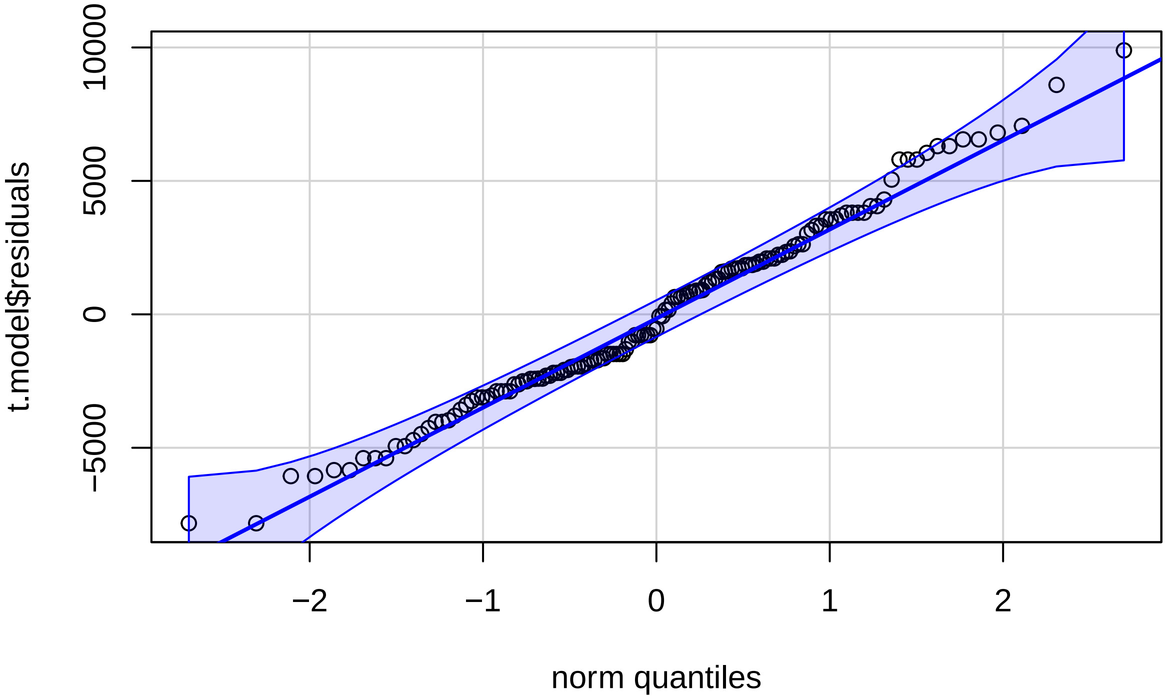
```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

	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
```

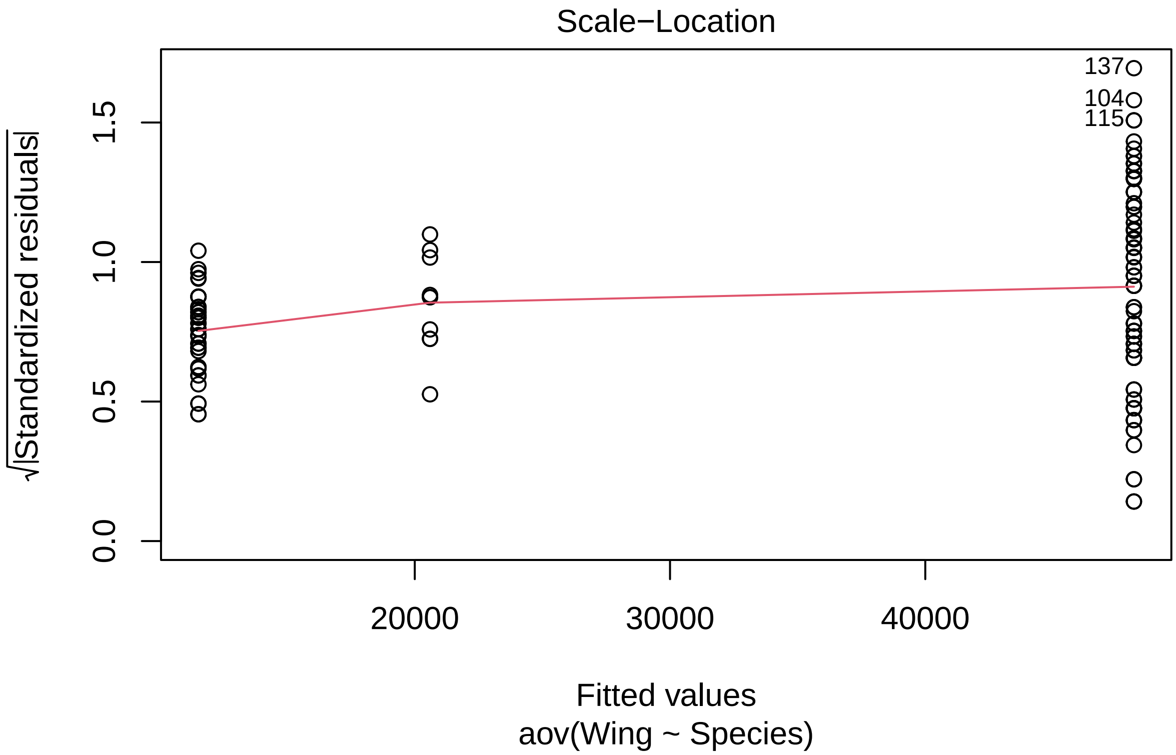
Get QQ Plot



Test for normality

```
##
##  Shapiro-Wilk normality test
##
## data:  t.model$residuals
## W = 0.98697, p-value = 0.1981
```

Plot variances



Test for const variance

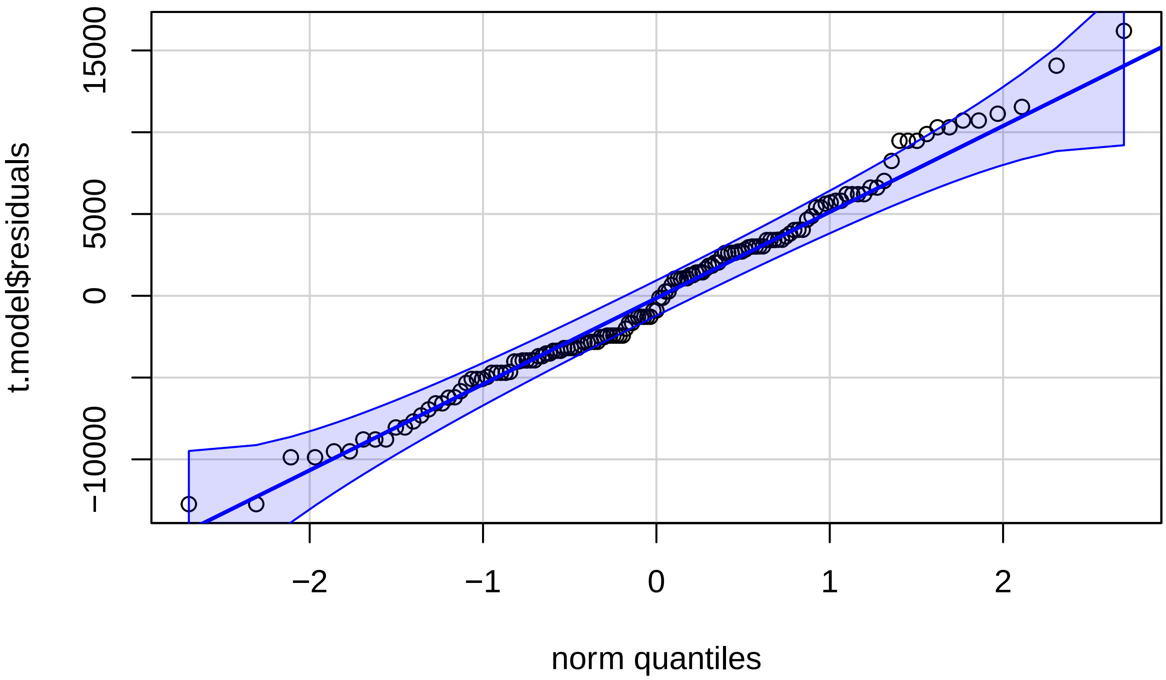
```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

	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
```

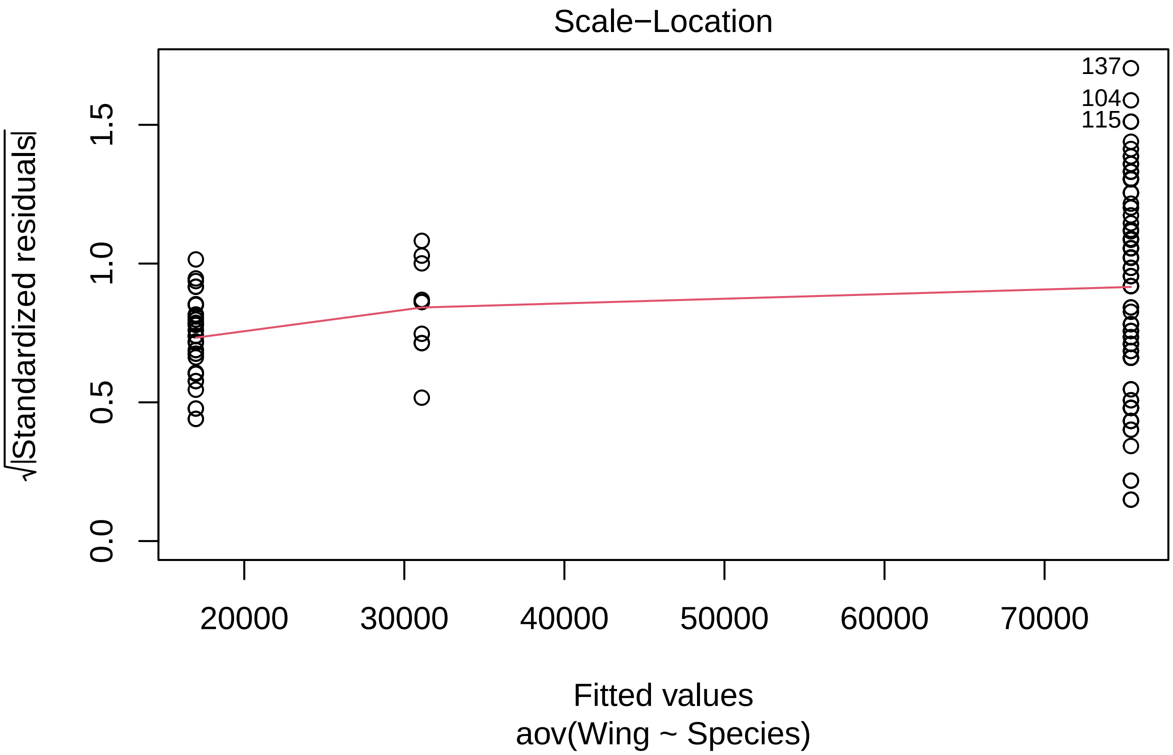
Get QQ Plot



Test for normality

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

Plot variances



Test for const variance

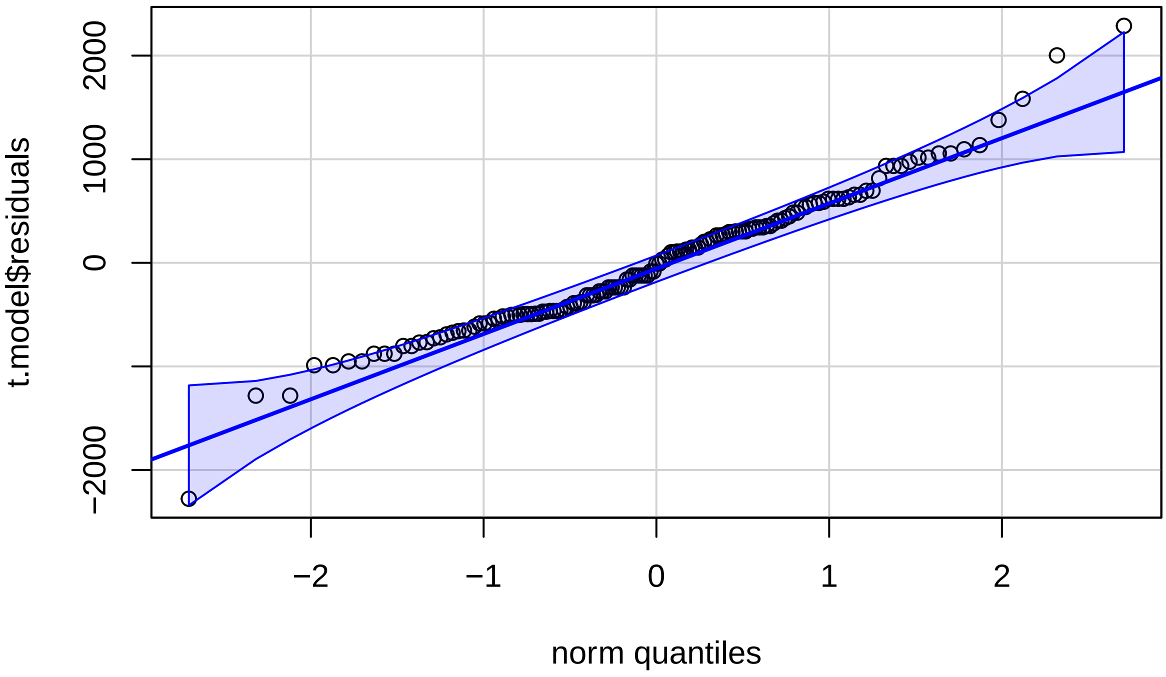
```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

	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
```

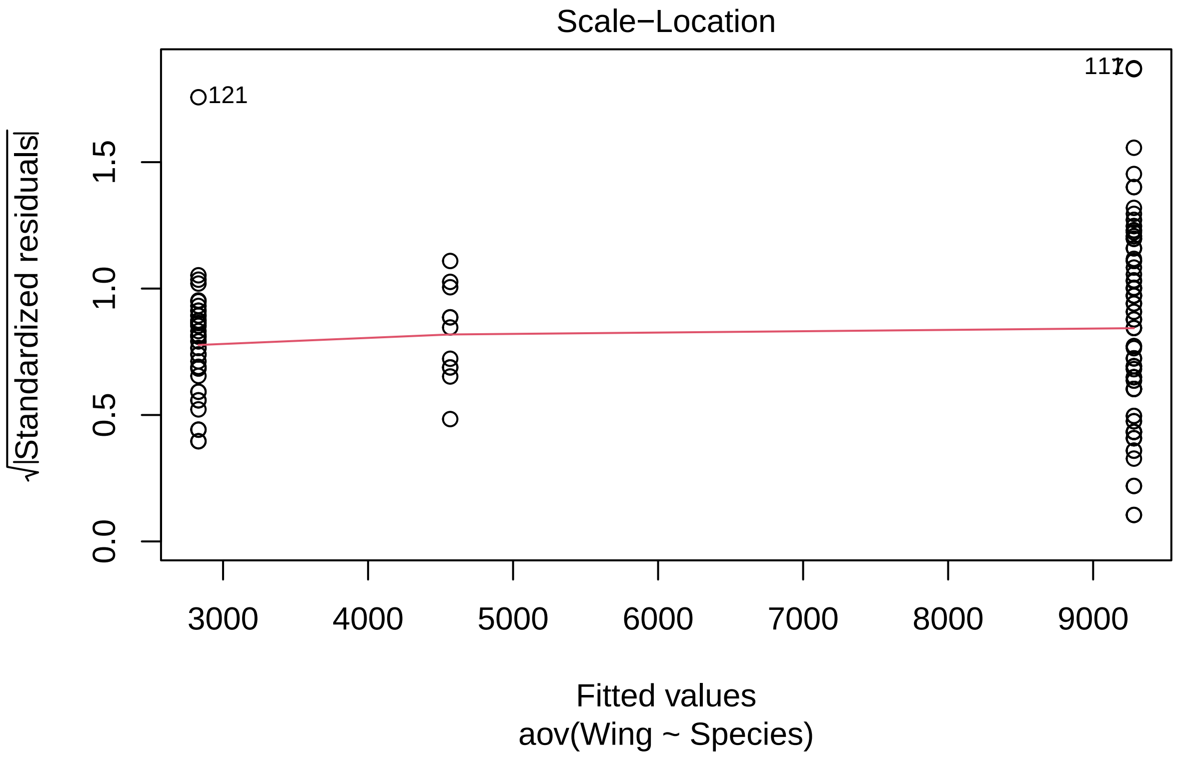
Get QQ Plot



Test for normality

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

Plot variances



Test for const variance

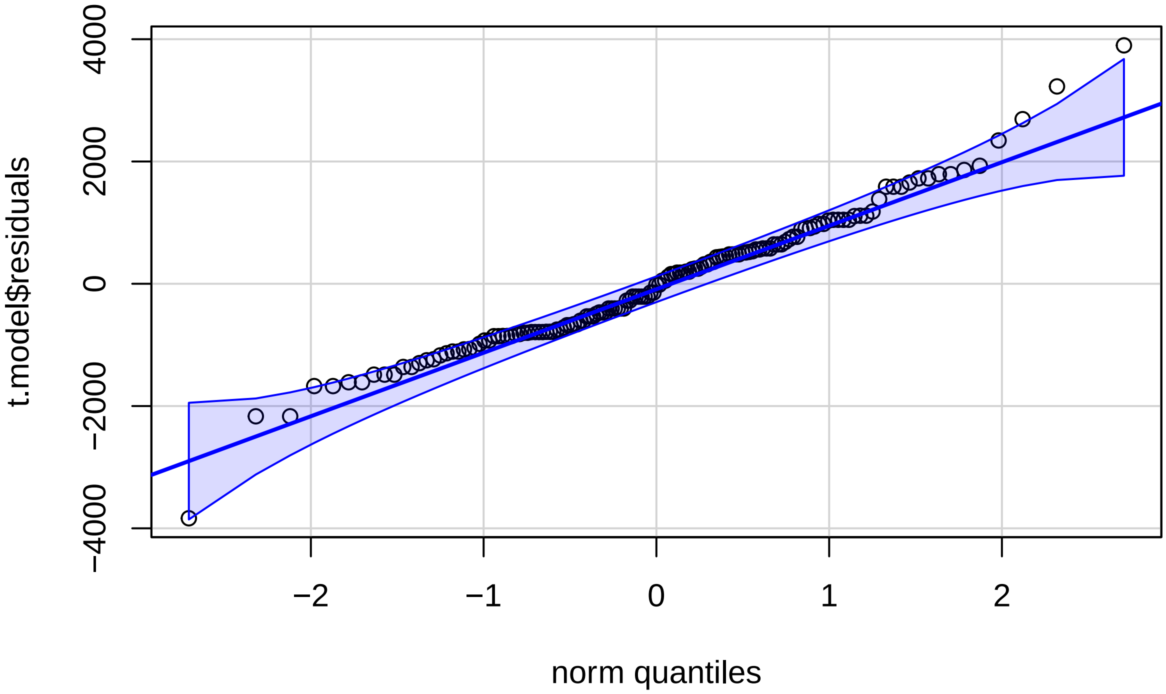
```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

	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
```

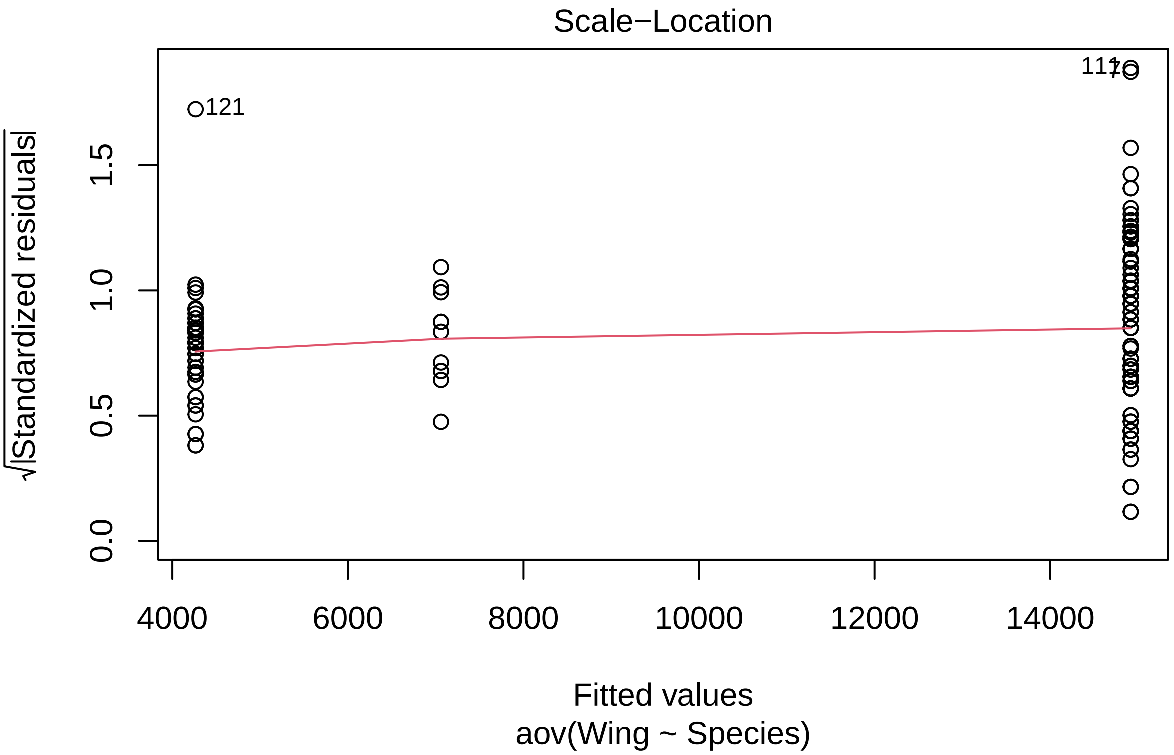
Get QQ Plot



Test for normality

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

Plot variances



Test for const variance

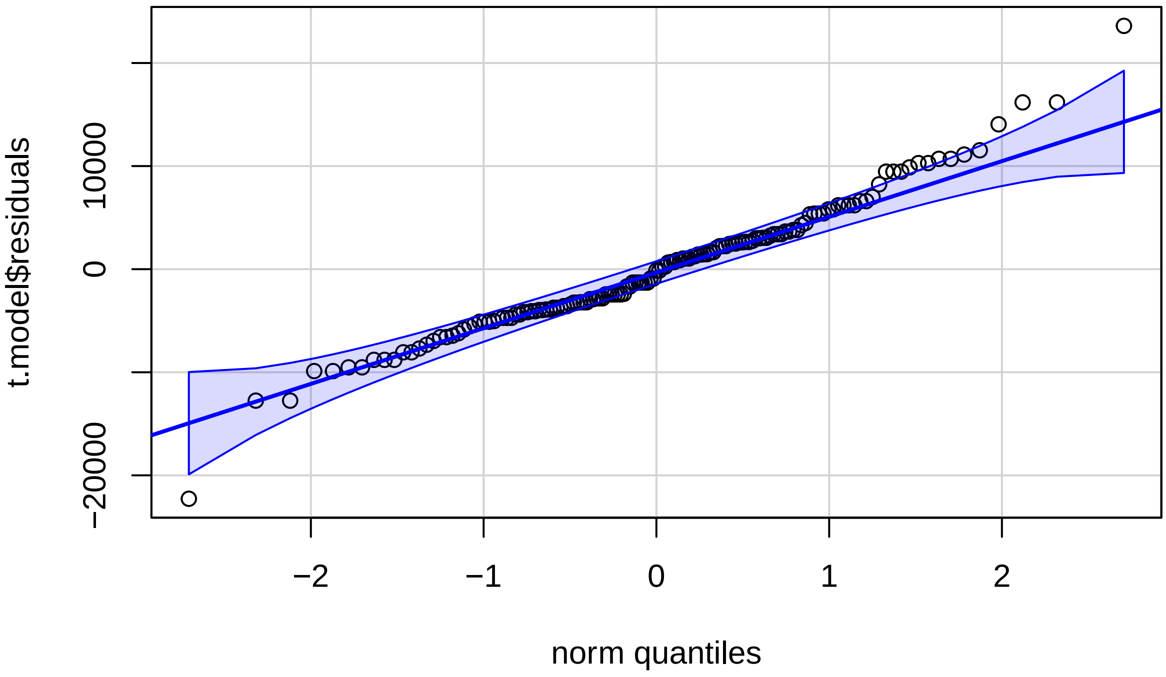
```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

	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
```

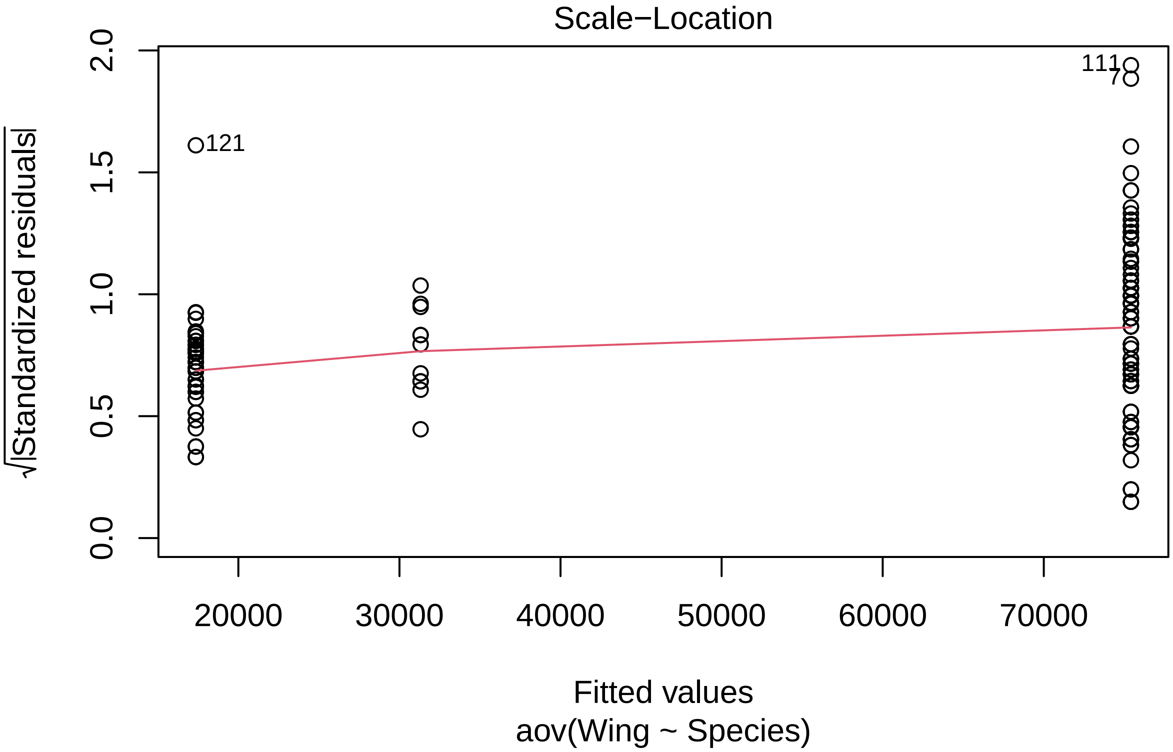
Get QQ Plot



Test for normality

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

Plot variances



Test for const variance

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

	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.